

1 **Sensory cheating: adversarial body**
2 **patterns can fool a convolutional**
3 **visual system during signaling**

4 Andres Laan^{1,*} and Gonzalo G. de Polavieja^{1,#}

5 May 19, 2018

6 ¹ *Champalimaud Neuroscience Programme, Champalimaud Center for the*
7 *Unknown, Lisbon, Portugal*

8
9 * andres.laan@neuro.fchampalimaud.org

10 # gonzalo.polavieja@neuro.fchampalimaud.org

11 **Abstract:**

12 Animals often assess each other by paying special attention to signals, which
13 help to communicate the quality of each individual. When there is a conflict
14 of interest between the signaler and the receiver, then the signaler has an
15 incentive to cheat by producing signals which exaggerate its apparent quality.
16 One opportunity for cheating might be to rely on sensory illusions, but it has
17 been difficult to study sensory cheating because we have lacked quantitative
18 models of complex visual perception. Here we address this problem by taking
19 advantage of recent advances in modeling visual brain areas as convolutional
20 neural networks. Given these models, we use the technique of adversarial
21 perturbations to show how sensory cheating can shape animal appearance
22 while nevertheless resulting in an evolutionarily stable signaling system. In
23 our simulations, animals typically evolve exaggerated color patterns which
24 might be analogous to the evolution of colorful body patterns in guppies.

25 **Introduction**

26 Convolutional neural networks (CNNs) have recently revolutionized the sci-
27 entific understanding of image processing and perception [1]. CNNs now

28 form the core component of most modern image recognition software and
29 are routinely used as data analysis tools across many domains. Unlike many
30 previous generations of machine learning models, CNNs are unique because
31 they consist of neuron-like elements and they may thus be viewed as candi-
32 date models for explaining the workings of biological visual systems as well.
33 Quantitative comparisons between neural recordings and CNNs have indeed
34 found a close resemblance between neural activity patterns inside CNNs and
35 the mammalian visual cortex [2, 3].

36 Improved quantitative models of visual perception may provide new ways
37 to theoretically analyze previously intractable problems. Here we use CNN
38 models to study the evolutionary stability of signaling in the presence of
39 conflicts of interest [4]. Our focus will be on the paradigmatic example of
40 aggressive signaling. During aggression, fighters display signals intended to
41 induce their opponent to surrender without a fight [5]. Typically, the individ-
42 ual who is of inferior fighting quality will be scared away by the higher quality
43 individual because the higher quality individual can afford to produce more
44 intense signals. Note that weak individuals theoretically have the option to
45 cheat by somehow producing a more intense signal than the stronger oppo-
46 nent. They are also motivated to do so because successful cheating would
47 lead to easy access to mates and resources. The puzzle of signaling is to
48 explain why cheating does not occur despite strong incentives to do so [4].

49 Classical models emphasize that such cheating cannot evolve if more in-
50 tense signals carry with them a greater cost of production, which only high
51 quality individuals are able to bare [6]. This is the standard argument in-
52 voked to explain phenomena like the large and uneconomical eyes of the
53 stalk-eyed flies or the peacock’s conspicuous tails[7].

54 However, the standard explanation can only be part of the answer, be-
55 cause it does not examine stability against sensory cheating. Sensory cheat-
56 ing entails a reshaping of the signal into a form which would make it appear
57 more intense than it really is to the senses of the receiver. The animal would
58 essentially use its own body as a canvas on which to craft a visual illusion.
59 For example, a courting animal might modify its body pigmentation pattern
60 to enhance its apparent height through the use of oriented vertical stripes [8]
61 and subsequently reap the benefits of the illusion through enhanced mating
62 success.

63 Many animals harness visual illusions in various contexts like camouflage,
64 escape and predator deterrence [9]. A paradigmatic case concerns animals
65 that display large false eyes in order to appear more threatening to predators
66 [10, 11]. Similar phenomena have been reported in the context of signaling
67 as well. Bower birds, for example, are known to actively shape the visual
68 environments of their mates to improve their own mating success [12, 13]. It

69 is therefore not inconceivable that in some species evolution might also shape
70 body patterns so as to trick the sensory systems of the receivers.

71 In order to quantitatively study the process of sensory cheating, we study
72 a signaling contest where the variable being estimated is body size. We
73 use body size as our variable because larger animals typically win aggressive
74 signaling contests and many animals actively display during conflict to signal
75 their size [5]. We train CNNs to estimate the sizes of model birds placed in
76 natural images. We then let the body pattern of the birds evolve in order
77 to fool the networks and we analyze the emergent dynamics to see if the
78 signaling system remains reliable throughout the process [14].

79 Results

80 Our study considers aggressive contests, where two individuals meet, assess
81 each other's size and the smaller individual subsequently retreats. Under
82 this scenario, any individual can improve its fitness if it can somehow modify
83 its appearance to appear larger than it really is to the perceptual system of
84 other animals.

85 To analyze this scenario, we first require a model of the size estimation
86 perceptual system. We therefore compiled a catalog of natural images
87 wherein we placed birds of various sizes. Then we trained a CNN to estimate
88 the size of the bird in each image. After that, we let the birds evolve their
89 appearance in ways that fooled the networks' perception.

90 We began by compiling a catalog of 4000 100 by 100 colored natural
91 images. The raw images were downloaded from the natural scene statistics
92 database [15] and 100 by 100 patches were extracted from the first 10 images
93 in the database. We created ten copies of each image and then we placed
94 inside these images the image of a model bird (**Figure 1** left panel). In
95 order to model natural variability in bird appearance, the bird varied in
96 height between 20 and 40 pixels, in rotation between -90 and 90 degrees and
97 its location in the image was also sampled randomly. Further, a different
98 sample of random noise was added to the body of the bird for each image
99 and its intensity was also varied. We thus created a highly variable and
100 non-trivially structured set of 40 000 images whose complexity was designed
101 to mimic the complexity of the natural environment. Sample images of the
102 resulting catalog can be seen in **Figure 1** left panel.

103 Next we trained a four-layered CNN to predict the size of the bird in
104 each image (see Supplementary methods). Training the CNNs by gradient
105 descent resulted in good predictive accuracy on both the training and the
106 test set (**Figure 1** right panels). The CNNs were thus able to solve the

107 task of separating the birds from the backgrounds and measuring the size of
108 the bird while ignoring irrelevant features like variation in orientation and
109 intensity.

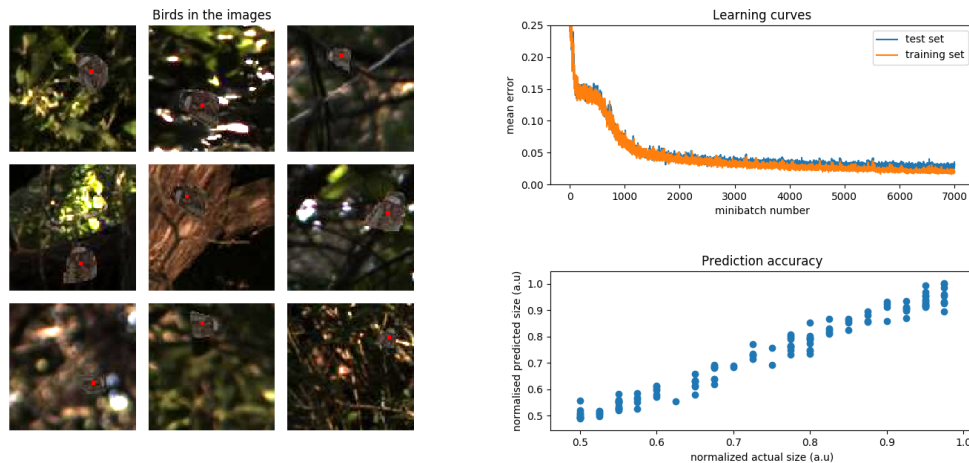


Figure 1: **Training the network.** Left panel: sample images from the catalog (the center of each bird is marked with a red dot for ease of viewing, the dots were not present in the images on which the CNN was trained). Top right panel: learning curves for the training and test set (batch size was 128 images). Bottom right: correspondence between the ground truth and the output of the trained network.

110 In order to model the evolution of body patterns, we adapted the tech-
111 nique used to find adversarial examples in artificial neural networks [16].
112 Briefly, we took three bird images of size 20, 30 and 40 pixels (small, medium,
113 large) and for each image, we calculated the gradient of the output of the
114 network with respect to each pixel of each image. This computation involved
115 estimating the average gradient by taking a sample mean across many back-
116 grounds, orientations and illumination levels (see **Supplementary meth-**
117 **ods**). Then we performed gradient ascent to make the birds appear progres-
118 sively larger over each iteration. The evolution of the largest bird’s appear-
119 ance can be seen in **Figure 2** on the left and the evolution of apparent size
120 is depicted in **Figure 2** right panel.

121 The birds increase in size by accentuating their edges and decreasing the
122 intensity of the center. They also evolve towards displaying unusual color
123 patterns which are not encountered in the training set. It is noted that
124 though the small, medium and large birds all considerably increase in appar-
125 ent size, the relative ranking of the sizes of the three birds remains stable

126 throughout evolution and signaling thus remains reliable [14]. Reliability may
127 be conserved because larger individuals are able to cheat more than smaller
128 individuals, because they have more body pixels which they can manipulate.
129 This may help larger individuals maintain their advantage over time.

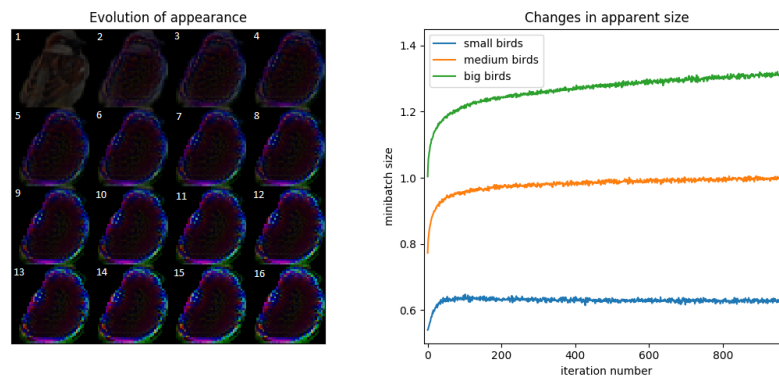


Figure 2: **Evolution of apparent size.** Left panel: changes in the appearance of the largest bird over time. Time is marked by numbers on the panels, each panel is separated from its predecessor by 60 iterations. Right panel: evolution of apparent size for the small, medium and large bird. Note that all birds considerably increase in size as time progresses but the relative ranking of the sizes nevertheless remains stable.

130 To establish the suitability of our methods for the study of biological
131 signaling, we further tested whether our conclusions were robust to variation.
132 In biological systems, the cheaters may need to be able to fool multiple
133 networks, since individual brains are known to vary [17]. Though most brains
134 are expected to produce similar outputs for similar inputs, they may be
135 achieving this feat in slightly different ways because internal connections will
136 vary somewhat due to factors like variability in brain development and early
137 visual environment. One way to simulate the variability would be to train
138 many different neural networks from different initial weight values and with
139 a different sequence of training examples.

140 We implemented this differential training process for five neural networks.
141 We then evolved a bird against the first network and then examined how the
142 findings generalized when the resulting body patterns were shown to the
143 four other networks. We found that examples developed against one network
144 typically generalized to the other four networks (**Figure 3** top left). We
145 further found that this conclusion held even if we changed the internal archi-
146 tecture of the network when we showed birds evolved against a network using
147 a relu non-linearity to a network using a hyperbolic tangent non-linearity as

148 shown in **Figure S1** left (hyperbolic tangent non-linearities could be viewed
149 as more biologically realistic due to its saturating behavior which more closely
150 mimics neuronal biophysics [18]). As expected, this conclusion did not hold
151 true when the images were shown to an untrained neural network (**Figure**
152 **S1** right).

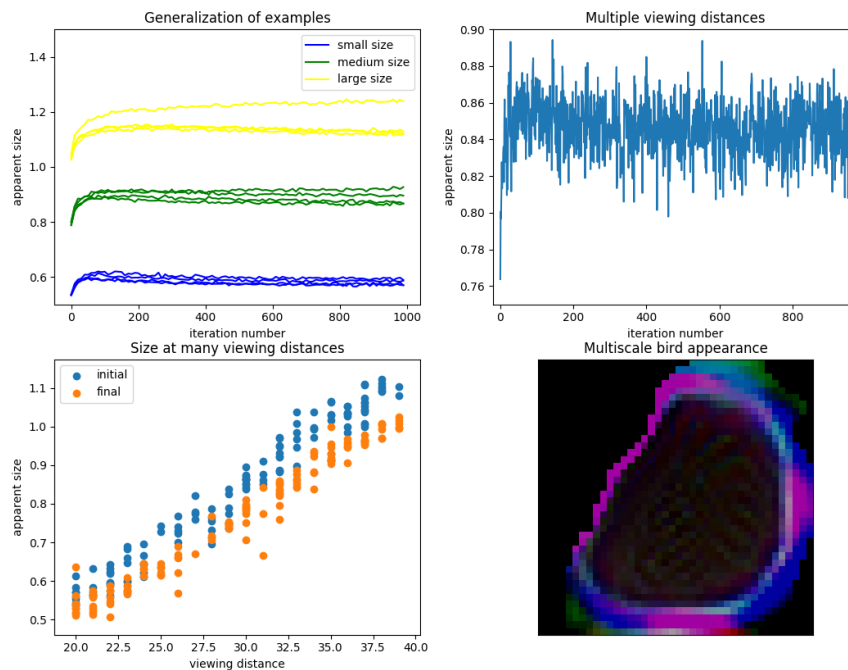


Figure 3: **Evolution of apparent size.** Top left: Birds evolved against one network are able to fool other networks which they did not encounter during evolution (each trace represents a separate network). Top right: evolution of mean apparent size when viewing distance varies. Bottom left: perceived size versus viewing distance at the beginning (orange) and end (blue) of evolution. The blue dots tend to lie above the red dots for all viewing distances indicating the ability of the mutant to robustly fool the CNN under many conditions. Bottom right: the final appearance of the large bird that fools CNNs at all distance.

153 We also tested whether our results are robust for all viewing distances.
154 When viewed from an identical distance, smaller animals should always oc-
155 cupy a smaller area on the retina than larger animals. However, signaling
156 displays are often complex spatial maneuvers during which the viewing dis-

157 tance may vary [19]. A good quality visual system would presumably be able
158 to distinguish between a big animal that is far away from a small animal that
159 is close even if both images occupy similar sizes on its retina. Based on these
160 concerns, we trained adversarial examples to be robust against variations in
161 viewing angle (see **Supplementary methods**). Our system was able to find
162 bird pigmentation perturbations which appeared larger than the original bird
163 image at all viewing distances (**Figure 3** top right and bottom panels). We
164 conclude that sensory cheating should be possible even against a visual sys-
165 tem which integrates information about image size with information about
166 inter-animal distances.

167 Discussion

168 We have demonstrated how convolutional neural networks could be applied
169 to the study of the evolutionary stability of signaling. We suggest that fu-
170 ture studies which examine the stability of signaling models should augment
171 traditional low-dimensional game theory analysis with a high-dimensional
172 analysis of signal form and natural image statistics [4, 6]. Our work shows
173 that unless sensory cheating is ruled out, the stability of any equilibrium
174 cannot be guaranteed.

175 Our approach made use of the technique of adversarial perturbations,
176 which was originally developed as a method to find small perturbations that
177 will cause machine classifiers to mis-classify an image [16]. Although these
178 perturbations were initially believed to be relevant only in the context of
179 artificial intelligence, recent research indicates that adversarial examples have
180 a limited ability to confuse human observers as well [20]. Our study indicates
181 that adversarial examples may also have further biological relevance in the
182 evolution of signaling and body patterns. Future work could also attempt
183 to apply these techniques to the study of segmentation systems and the
184 evolution of camouflage [21].

185 Our work is not the first to recognize the usefulness of explicit cognitive
186 models for the study of evolution. Pioneering theoretical work by Enquist
187 and others used artificial neural networks like the multi-layer perceptron as
188 a tool in the theoretical study of evolution [22]. This early work was of
189 limited applicability because slow computers did not allow these systems
190 to be trained on complex real world tasks. With the availability of fast
191 modern hardware, it should become increasingly easy to design and probe
192 the function of complex pattern recognition systems through an evolutionary
193 lens.

194 One of the empirical findings of our work was that in the later stages

195 of evolution, the model birds evolved to display unusual colors. This is an
196 outcome that likely occurs because adaptive systems are typically tuned to
197 work accurately only on their training domain as they do not face selective
198 pressure to correctly analyze out of domain signals. Since bright and pure
199 colors lie outside the typical statistics of natural images, it is not surprising
200 that these signals turned out to be effective at driving spurious signaling
201 activity in the networks. These findings may have some parallels with the
202 evolution of bright body colors in Trinidadian guppies [23]. When relived
203 from predation pressure, Trinidadian guppies evolve to display bright colors
204 for the purposes of increasing their attractiveness to potential mates. It may
205 be the case that these bright body patterns function partly as adversarial ex-
206 amples or hyper-stimuli that are particularly effective at driving the activity
207 of the sexual quality assessment network of the fish brain.

208 Our work focused on the evolution of body patterns without considering
209 the simultaneous evolution of the neural network used for assessment. We
210 made this modeling choice because signaling equilibriums may be understood
211 as Bourgeois strategies (where the asymmetry happens to be correlated but
212 need not remain so throughout evolution) and no individual has an incentive
213 to deviate from consensus assessments [24]. Since our modeling finds that
214 signaling remains reliable, it could also serve as a useful model for scenar-
215 ios where body pattern evolution is for some reason much more rapid than
216 the evolution of the assessment network. For more complex scenarios like
217 the study of sexual selection, this approximation may not remain valid and
218 future work must find ways to extend our methods to take into account the
219 aforementioned complexities [14].

220 Finally, it will be interesting to study if certain body patterns or brain
221 architectures are less vulnerable to cheating. It might be expected that pure
222 bright colors which are already unusual for a given environment and easy to
223 separate from the background might be rather immune to cheating. Also,
224 there may be other neural networks which utilize movement information or do
225 a more complex segmentation that will prove more difficult to hack. Future
226 work will need to explore these issues in more extensive detail.

227 **Supplementary materials**

228 **Supplementary figures**

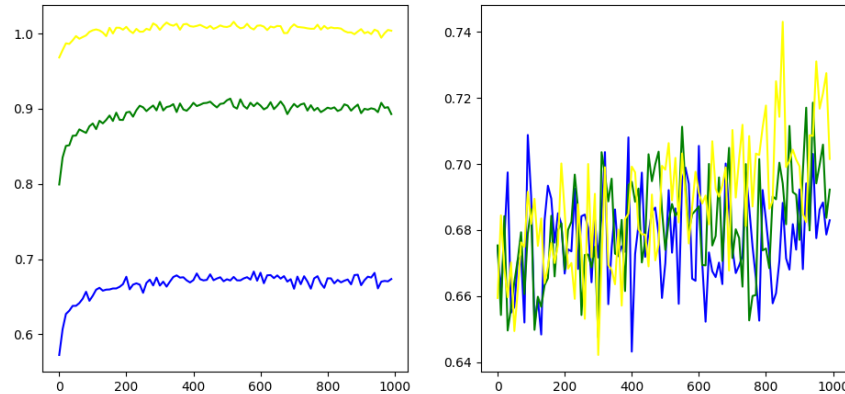


Figure 4: **Generalization of examples.** Left: Birds evolved against a relu network shown to a hyperbolic tangent network. Right: Birds evolved against a trained relu network shown to an incompletely trained (100 instead of 7000 mini-batches) relu network.

229 **Supplementary methods**

230 **Details on neural networks**

231 We used 4 convolutional layers with relu non-linearities, each followed by a
232 2-by-2 max pooling layer. All layers used 5x5 filters. Filter numbers by layer
233 were 32, 64, 64, 64. The fully connected layer used 512 neurons. We trained
234 the network using gradient descent on the mean squared error loss function
235 with the Adam optimizer using a learning rate of 10^{-4} with mini-batches of
236 size 128. The training set consisted of 35 000 images from which mini-batches
237 were sampled randomly. For the relu networks training process used 7000
238 mini-batches. For the tanh non-linearity training took 200 000 mini-batches.
239 Training was implemented in Tensorflow.

240 **Adversarial examples**

241 During evolution, birds will evolve towards greater apparent size. The image
242 of a bird can be regarded as a set of pixels. To predict how the birds will

243 evolve, we must predict which pixel changes would increase the expected
244 apparent size of the bird. In other words, we must predict the gradient of
245 the expected apparent size with respect to each pixel of the bird. In order
246 to estimate the expected value of the gradient, we must average over all
247 possible locations, orientations, backgrounds, noise perturbations, etc. We
248 calculate the estimate using Monte Carlo sampling. We first embed the bird
249 in 128 images, whose orientation, location, background, etc statistics are
250 sampled from the same distribution as was used for generating the training
251 set. For each image, we use standard Tensorflow procedures to estimate the
252 gradient of the output (the estimated size) with respect to all the image
253 pixels. Then we back-transform this gradient onto the bird image template
254 by shifting and rotating the image such that the birds in all the images will
255 line up exactly. The gradient estimate is the sample mean of these back-
256 transformed gradients. Then we add the learning-rate weighted gradient
257 onto the bird images to obtain a new bird image and repeat the procedure
258 again. The same procedure was used to in the viewing distance invariant
259 scenario, but there we also added an extra image scaling step to the back-
260 transformation step to compensate for the fact that the size of the bird in
261 each images varied depending on the viewing distance (the viewing distance
262 was sampled uniformly at random between 20 and 40 units). Further details
263 on the code are available from the authors on request and all code will be
264 deposited at a public repository after publication of the manuscript.

265 References

- 266 [1] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with
267 deep convolutional neural networks. In: Advances in neural information
268 processing systems; 2012. p. 1097–1105.
- 269 [2] Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsuper-
270 vised, models may explain IT cortical representation. PLoS computa-
271 tional biology. 2014;10(11):e1003915.
- 272 [3] Kriegeskorte N. Deep neural networks: a new framework for modeling
273 biological vision and brain information processing. Annual review of
274 vision science. 2015;1:417–446.
- 275 [4] Smith JM, Harper D. Animal signals. Oxford University Press; 2003.
- 276 [5] Arnott G, Elwood RW. Assessment of fighting ability in animal contests.
277 Animal Behaviour. 2009;77(5):991–1004.

- 278 [6] Grafen A. Biological signals as handicaps. *Journal of theoretical biology*.
279 1990;144(4):517–546.
- 280 [7] Emlen DJ. *Animal weapons: the evolution of battle*. Henry Holt and
281 Company; 2014.
- 282 [8] Thompson P, Mikellidou K. Applying the Helmholtz illusion to fash-
283 ion: Horizontal stripes won't make you look fatter. *i-Perception*.
284 2011;2(1):69–76.
- 285 [9] Kelley LA, Kelley JL. Animal visual illusion and confusion: the impor-
286 tance of a perceptual perspective. *Behavioral Ecology*. 2013;25(3):450–
287 463.
- 288 [10] De Bona S, Valkonen JK, López-Sepulcre A, Mappes J. Predator
289 mimicry, not conspicuousness, explains the efficacy of butterfly eyespots.
290 *Proc R Soc B*. 2015;282(1806):20150202.
- 291 [11] Stevens M, Hardman CJ, Stubbins CL. Conspicuousness, not eye
292 mimicry, makes eyespots effective antipredator signals. *Behavioral Ecol-*
293 *ogy*. 2008;19(3):525–531.
- 294 [12] Endler JA, Endler LC, Doerr NR. Great bowerbirds create theaters
295 with forced perspective when seen by their audience. *Current Biology*.
296 2010;20(18):1679–1684.
- 297 [13] Kelley LA, Endler JA. Illusions promote mating success in great bower-
298 birds. *Science*. 2012;335(6066):335–338.
- 299 [14] Searcy WA, Nowicki S. *The evolution of animal communication: reli-*
300 *ability and deception in signaling systems*. Princeton University Press;
301 2005.
- 302 [15] Geisler WS, Perry JS. Statistics for optimal point prediction in natural
303 images. *Journal of Vision*. 2011;11(12):14–14.
- 304 [16] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow
305 I, et al. Intriguing properties of neural networks. arXiv preprint
306 arXiv:13126199. 2013;.
- 307 [17] Maguire EA, Gadian DG, Johnsrude IS, Good CD, Ashburner J, Frack-
308 owiak RS, et al. Navigation-related structural change in the hippocampi
309 of taxi drivers. *Proceedings of the National Academy of Sciences*.
310 2000;97(8):4398–4403.

- 311 [18] Byrne JH, Heidelberger R, Waxham MN. From molecules to networks:
312 an introduction to cellular and molecular neuroscience. Academic Press;
313 2014.
- 314 [19] Laan A, Iglesias M, de Polavieja G. Opponent assessment and conflict
315 resolution through mutual motor coordination. bioRxiv. 2017;p. 208918.
- 316 [20] Elsayed GF, Shankar S, Cheung B, Papernot N, Kurakin A, Goodfellow
317 I, et al. Adversarial Examples that Fool both Human and Computer
318 Vision. arXiv preprint arXiv:180208195. 2018;.
- 319 [21] Stevens M, Merilaita S. Animal camouflage: current issues and new per-
320 spectives. *Philosophical Transactions of the Royal Society B: Biological*
321 *Sciences*. 2009;364(1516):423–427.
- 322 [22] Enquist M, Ghirlanda S. *Neural networks and animal behavior*. Prince-
323 ton University Press; 2013.
- 324 [23] Endler JA. Natural selection on color patterns in *Poecilia reticulata*.
325 *Evolution*. 1980;34(1):76–91.
- 326 [24] Sherratt TN, Mesterton-Gibbons M. The evolution of respect for prop-
327 erty. *Journal of evolutionary biology*. 2015;28(6):1185–1202.