Testing the equivalency of human "predators" and deep neural networks in the detection of cryptic moths

Mónica Arias, Lis Behrendt, Lyn Dressler, Adelina Raka, Charles Perrier, Marianne Elias, Doris Gomez, Julien Renoult, Cynthia Tedore

Testing the equivalency of human "predators" and deep neural networks in the detection of cryptic moths

Authors
Mónica Arias[1,*], Lis Behrendt[2], Lyn Dreßler[2], Adelina Raka[2], Charles Perrier[3], Marianne Elias[4], Doris Gomez[5,], Julien P. Renoult[5], Cynthia Tedore[2,5,*]

Affiliations:
[1] PHIM, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France. Orcid number: 0000-0003-1331-2604
[2] Univ. Hamburg, Faculty of Mathematics, Informatics and Natural Sciences, Institute of Zoology, Hamburg, Germany
[3] CBGP, INRAe, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, Montpellier, France
[4] ISYEB, CNRS, MNHN, Sorbonne Univ., EPHE, Univ. Antilles, Paris, France
[5] CEFE, CNRS, Univ. Montpellier, Univ. Paul Valéry Montpellier 3, EPHE, IRD, Montpellier, France

* **Corresponding authors**
Corresponding author emails: monica.arias@cirad.fr, cynthia.tedore@uni-hamburg.de

Abstract
Researchers have shown growing interest in using deep neural networks (DNNs) to efficiently test the effects of perceptual processes on the evolution of color patterns and morphologies. Whether this is a valid approach remains unclear, as it is unknown whether the relative detectability of ecologically relevant stimuli to DNNs actually matches that of biological neural networks. To test this, we compare image classification performance by humans and six DNNs (AlexNet, VGG-16, VGG-19, ResNet-18, SqueezeNet, and GoogLeNet) trained to detect artificial moths on tree trunks. Moths varied in their degree of crypsis, conferred by different sizes and spatial configurations of transparent wing elements. Like humans, four of six DNN architectures found moths with larger transparent elements harder to detect. However, humans and only one DNN architecture (GoogLeNet) found moths with transparent elements touching one side of the moth's outline harder to detect than moths with untouched outlines. When moths took up a smaller proportion of the image (i.e., were viewed from further away), the camouflaging effect of transparent elements touching the moth's outline was reduced for DNNs but enhanced for humans. Viewing distance can thus interact with camouflage type in opposing directions in humans and DNNs, which warrants a deeper investigation of viewing distance/size interactions with a broader range of stimuli. Overall, our results suggest that humans and DNN responses had some similarities, but not enough to justify widespread use of DNNs for studies of camouflage.

Keywords
deep learning, human observers, artificial prey, background matching, disruptive coloration, transparency.

Introduction

Artificial intelligence algorithms, and deep neural networks (DNNs) in particular, are becoming common tools for tackling questions in ecology and evolution while reducing human labour (Christin et al., 2019; Norouzzadeh et al., 2018; Tosh & Ruxton, 2010). DNNs consist of two parts, an encoder that projects a high-dimensional input, e.g., an image of a visual pattern or an entire animal, into a lower dimensional space, and a classification or regression function that uses encodings of the low dimensional space to make predictions (Charpentier et al., 2020; J. F. H. Cuthill et al., 2019; Wham et al., 2019). DNNs are made up of multiple layers of neurons that lack interconnections within layers but are connected to neurons in adjacent layers. The firing intensity of a neuron depends on the strength of signal received by connected upstream neurons and the neuron's nonlinear activation function. Learning comes about by iteratively tweaking the weight of each neuron's output to downstream neurons until a more accurate final output is attained. Although the architecture and computational mechanics of artificial neural networks were originally inspired by biological neural networks, the primary goal of deep neural networks today is to produce accurate outputs (in our case, categorizations) rather than to mimic the mechanics of biological neural networks. This could potentially be problematic for studies of camouflage, since prey color patterns evolve in response to the perceptual and cognitive processes of their predators (Merilaita et al., 2017), not of DNNs.

Despite fundamental differences between the inner workings of biological and artificial neural networks, the spatial distribution of images of different classes of objects in the representational space of DNNs has been shown in various studies to be correlated with that of the primate visual cortex (Cadieu et al., 2014; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Mély & Serre, 2017; Renoult et al., 2019; Yamins et al., 2014). Distances between images of patterns or entire animals in the representational space of the DNN encoder have also been shown to be correlated with their perceived visual similarity in humans (Zhang et al., 2018). And a recent study found that simulated moth-like prey that have evolved to be cryptic to DNNs become increasingly cryptic to humans as their evolution proceeds (Talas et al., 2020). That said, whether this works in the reverse direction, i.e., that increasingly cryptic prey to humans (or other animals) are also increasingly cryptic to DNNs, is unknown. This deserves testing and is the goal of the present study.An easy way to implement DNNs in ecology and evolution research is to take a convolutional DNN that has been pre-trained by computer scientists to recognize many different types of object classes and then retrain it to recognize a new stimulus or set of stimuli. . This procedure is called transfer learning, and is convenient because it requires a much smaller training set than would be required for a DNN with no previous training. In camouflage research, the new task is often a binary discrimination task, i.e., detecting the presence/absence of a specific object, such as a prey item. Unfortunately, DNNs widely available for transfer learning only accept three-channel images ("RGB" for red, green, and blue), excluding the possibility to include more than three spectral bands. This may be problematic when using DNNs to mimic animal visual systems very different from our own, especially those with more than three spectral channels, as are common in many animals, from birds to butterflies. However, in several studies, humans with their three-channel vision have shown reactions similar to those of birds towards images of colorful objects. For example, birds and humans can use aggregation of conspicuous prey as an unprofitability signal (Beatty et al., 2005; Finkbeiner et al., 2012), and blue tits (Kazemi et al., 2014) and humans (Sherratt et al., 2015) focus more on salient color cues than on pattern or shape information to classify prey as unpalatable. Reactions of humans and birds are also similar towards cryptic prey. For example,

disruptive coloration (i.e. contrasting elements touching prey contour) is more successful at concealing prey in comparison to background matching to the eyes of both humans (Troscianko et al., 2013) and birds (I. C. Cuthill et al., 2005). Similarly, both humans (Tucker & Allen, 1988) and birds (Bond & Kamil, 1998) more often overlook rare than common cryptic prey (apostatic selection).

Here we make a first attempt at testing whether widely-available DNNs, trained through transfer learning to a novel task, mimic the human perceptual system well enough to test hypotheses about the relative detectability of different cryptic prey to, at the very least, the perceptual system they are designed to mimic, i.e., humans. While one could debate how similarly DNNs should perform to humans in order to be considered sufficiently accurate for ecological research, we take, as a bare minimum criterion, that the directionality of effects should be the same. Specifically, we exposed both DNNs and humans to RGB pictures of five different cryptically-colored artificial moth morphs placed on tree trunks. These moths varied in their degree of crypsis by exhibiting transparent elements that varied in size and position (i.e., transparent windows touching or not touching prey outlines). The opaque portions of these moths appear cryptically colored to humans and exhibit low color and achromatic contrasts to birds as well (Arias et al., 2020). We chose these stimuli because they were recently used in an experiment testing the effect of size and position of transparent elements on prey detectability to wild birds in the field (Arias et al., 2021). By using the same stimuli as those used in the previous study, we gain some insight into the suitability of using DNNs and/or humans as surrogate predators for birds.

Material and methods

Artificial prey

We took pictures of artificial grey moths identical to the moths used in the field experiment from Arias et al (2021). Prey were designed to represent different crypsis levels. Therefore, they constitute stimuli associated to different degrees of both background matching and disruptive coloration, camouflage strategies that have been proven differently performant (I. C. Cuthill et al., 2005; Fraser et al., 2007). The same pictures were used in both the DNN and the human experiments. We tested five types of artificial grey moths with different wing visual characteristics (Figure 1): 1) opaque (O morph), 2) with small transparent windows occupying 15% of the total wing surface (SW morph), 3) with large transparent windows occupying 46.6% of the total wing surface and not touching any edge of the wing (LW morph), 4) with large transparent windows touching the bottom edge of the wing (BE morph), and 5) with large transparent windows touching all three wing edges (as each window is touching two edges of the three moth edges, B3E morph). In the latter three morphs, the total surface area of the opaque portions of the wings were identical. The opaque parts of the wings were made out of paper. For morphs including transparency, windows were cut out from the paper moth and a transparent film was added underneath the remaining parts. On top of the moth wings, we added an artificial grey body matching the wing colour. A tablet (iPad mini 2) and a smartphone (Moto G5, Android 8.1.0) using the "U camera" application to produce square-shaped photos were used to take pictures of artificial prey. No HDR option was activated on either of the cameras. The device used was mounted on a tripod to take pictures of each of five artificial moth morphs, each pinned to the same position on a tree trunk, as well as the tree trunk alone with no moth present. This controlled setup ensured that we did not unintentionally photograph certain morphs against more challenging backgrounds. Each morph was photographed against 222 unique visual backgrounds.

Deep neural networks experiment

We trained deep neural networks (DNNs) to detect all five moth morphs on tree trunks and then compared their responses to those of human "predators" when detecting the same prey.

Selected DNNs and general training

Rather than training DNNs from scratch (i.e., initializing all neuronal outputs with random weights), we used convolutional DNNs pre-trained to correctly identify 1000 object categories from more than a million images from the ImageNet database. Pre-trained networks have already learned to detect image features common to many classes of objects, and have pre-defined neuronal weights that are useful for learning new combinations of features. We took such DNNs as a starting point, and then used transfer learning to train them on the new task of detecting artificial moths. In transfer learning, the last few layers of an existing DNN are typically replaced with new ones that rapidly adapt to a new classification task. Transfer learning allows a neural network to reach high performance in a classification task even with a relatively small training dataset. To test the robustness of results obtained from DNNs, we used 100 replicates of each of six different DNNs that were similarly trained: AlexNet, VGG-16, VGG-19, ResNet-18, SqueezeNet, and GoogLeNet.

Out of the 222 images of each morph, the training set included 144 randomly-selected images of each of the five morphs and each corresponding background image in quintuplicate. We included each background in quintuplicate so that there would be the same number of positive (with a moth) and negative stimuli (without a moth) in the training set. Thus, in total, 1,440 images were used in training. Thirty-nine randomly-selected images of each morph and their corresponding background images in quintuplicate were used in the validation set (for a total of 390 images), and the remaining 39 images (out of 222) of each morph and the background were used in the test set (for a total of 234 images).

During training, weights were frozen in the first 25% of layers to preserve feature detectors useful for recognizing many types of objects already learned from the ImageNet database (this amounted to two layers in AlexNet, four layers in VGG-16, and five layers in SqueezeNet, ResNet-18, VGG-19, and GoogLeNet). The initial learning rate in subsequent layers was set to $2*10^{-4}$, and was increased by a factor of ten in the final, fully-connected layer. The learning rate across all layers was reduced by a factor of 0.1 every ten epochs. The L_2 regularization term was set to 0.0001. Different solvers, which identify optimal weights, were used for different DNNs to maximize the learning capability (measured as the test set accuracy) of the different DNN architectures. The adaptive moment estimation (Adam) solver was used for AlexNet, ResNet-18, SqueezeNet, and GoogLeNet, and the stochastic gradient descent with momentum (sgdm) solver was used for VGG-16 and VGG-19. Minibatch size (the subset of images used during one iteration of training) was set to 80. Each minibatch contained eight sets of ten images. Each set of ten images had the same background, five of which contained one each of the O, SW, LW, BE, and B3E morphs. The remaining five in each set of ten was the same background image in quintuplicate.

Training was halted and considered complete when the loss (a measurement of DNN uncertainty) on the validation set was higher than the previously smallest loss three times, in order to prevent overfitting (Ying, 2019). The final output network was the one with the lowest validation loss. We repeated this training and testing procedure 100 times, each time retraining the same initial DNN (pretrained on the ImageNet database) and randomly selecting which images (out of the full 222 images available per morph) were to be allocated for training, for validation, and for testing.

After training and testing, each DNN returned a score between 0 and 1 indicating its level of confidence that there was a moth in each picture. This score was determined by a sigmoid binary classifier (softmax layer in the DNN). Images with scores greater than 0.5 were classified as displaying a moth. To verify that DNNs were detecting moths using features of the wings, and not the body alone, we generated class activation maps (CAMs) of the final layer of the DNNs. These maps indicate which parts of the image contribute the most towards determining that the image contained a moth. SqueezeNet, ResNet-18, and GoogLeNet can produce such maps, but AlexNet, VGG-16, and VGG-19 cannot due to the fact that their final layer consists of multiple interconnected layers. All DNN training and testing and the generation of class activation maps were conducted using MATLAB (The MathWorks In., 2019).

Analyses of DNN results

We compared DNN architecture performances at moth detection by comparing their false positive rates (proportion of background images where DNNs had a score higher than 0.5 while there was no moth). Secondly, we compared DNN differences in detection between morphs. We created a binary response variable that only included responses to pictures with moths in them, and that was set to one for DNN scores higher than 0.5 (moth detected) and set to zero for DNN scores lower than 0.5 (moth not detected). Then, we fitted generalised linear mixed models (GLMM) for each DNN architecture assuming a binomial distribution, using the binary variable as a response variable. The explanatory variables were moth size in the picture (an inverse proxy for viewing distance; measured as the diameter of the moth's broadest dimension and rescaled to have the minimum value at 0 and the maximum at 1) and moth morph. Different models were fitted including or not interactions between the explanatory variables. The AICc criterion was used to select the model that best fit our data. Moth morph had five levels: O, SW, LW, BE and B3E. We tested the following contrasts: a) O was more detected than other morphs, b) morphs with small windows (SW) were more detected than morphs with large windows (LW), c) morphs with uninterrupted edges (LW) were more detected than morphs with one broken edge (BE) and d) morphs with only one broken edge BE) were more detected the morphs with three broken edges (B3E). Random effects included picture set (pictures sharing the same background) and replicates (as each pre-trained DNN was trained to detect the experimental stimuli 100 independent times).

Human experiment

Before the game

We developed an online game using the Testable platform (https://www.testable.org/). The game was only available for computers, not for smartphones or tablets. As it was an online game, we made it as standardized as possible for all participants by asking them to follow several simple steps before the test. Participants were asked to hold up a credit card to a bar on the screen and adjust the size of the bar on the screen using arrow keys until it was the same size as the credit card. This enabled us to present the photos in equal sizes (a 4.5 cm square) to all participants. The small size at which the images were presented was necessary in order for the images to have the same resolution as that required by the DNNs while not appearing pixelated. Participants were asked to sit one arm-length distance away from their computer screen in order to ensure a consistent viewing distance across images seen by a given participant. Although people's arm lengths vary, this effect was controlled for by including participant height and ID in the later statistical modelling. People always wearing glasses were asked to keep their glasses on. Participants were requested to concentrate on the fixation cross between the trials and to stay

focused and concentrated throughout the whole test and to keep interruptions at a minimum level. They were asked to look carefully at the photos to minimize false negatives. Participants were instructed to press the right arrow key when they spotted a moth and the left arrow key if they found no moth, and, if they did see a moth, to press the right arrow key as quickly as possible. Instructions were given in easy-to-read text in both German and English languages. Then, a short learning section was conducted, including one photo of each morph and one background photo. These photos were excluded from the real test. This training section ensured that participants had been previously exposed to all morphs, understood what they were looking for, and were familiar with the mechanics of the game. To start the real test, participants had to press 'enter' to ensure that they only started the test when they were focused and interruptions and distractions were set to a minimum.

During the game

For each participant, a subset of 110 pictures was randomly picked out of the entire pool of 1332 pictures used in the DNN experiment. This subset always included 11 photos of each moth morph (in total 55) and 55 photos corresponding to background images containing no moth. In a maximum of five seconds, participants had to choose whether there was a moth in the picture or not. We recorded the time a participant needed to press the key (reaction time, RT), and we assumed that short RTs corresponded to moths that were easier to detect. Additionally, the number of correct choices was counted. If the user did not respond within the five second picture presentation period, their response was recorded as correct if there was no moth in the picture and incorrect if there was a moth in the picture. To ensure participants were engaged and present at their computer throughout the experiment, if no response was recorded within the five second response time, a keystroke was required in order to proceed to the next trial. Between two pictures, a fixation cross was shown in the center of the screen to generate a specific starting point for all participants. The inter-trial interval (ITI) varied between 300 to 1500 milliseconds to prevent participants from getting used to the ITI, possibly affecting RTs. After 55 photos in random order, a break was included showing a black screen saying "short break" for five seconds. This break was intended to provide a short relaxation for the eyes and to enhance concentration. After that, another set of 55 pictures was shown.

Data collected from each participant included: sequence of the pictures, RT for each picture, choice correctness, age, sex, height, and participant code (to keep track of the performance of each participant in an anonymous way). In total, data from 283 participants were collected.

Analyses of the human experiment results

As for the DNNs, we first compared participants' performances at moth detection by comparing their false positive rates (proportion of background images where participants said there was a moth while there was none). Then, we analysed participants' performance in response to the different morphs by analysing morph survival. We fit mixed effect Cox Proportional Hazards models (Cox, 1972) using the *coxme* package (Therneau, 2020) in R (R Foundation for, 2014). Morph survival was calculated using both the time that participants took to find the moths (RT) and the numbers of each morph that were found per participant (correct answers when a moth was in the image). Moths that were not found by participants were considered as censored data in the model. Moth morph, moth size in the picture, participant height (expected to change the perceived size of the moths by modifying the distance to the screen) and position in the sequence of pictures shown to each participant (sequence) were included as fixed effects in the

model. Different models were fitted including different interactions between the explanatory variables. The AICc criterion was used to select the model that fit the data best. For morphs, we tested the same contrasts as in the DNN experiment: a) O was more detected than other morphs, b) morphs with small windows (SW) were more detected than morphs with large windows (LW), c) morphs with complete windows (LW) were more detected than morphs with one broken edge (BE) and d) morphs with one broken edge (BE) were more detected than morphs with three (B3E) broken edges. Picture set (out of 222) and participant sex, age, and code were included as random effects. Finally, we visualized the predictability of human responses by DNN outputs by fitting linear models using proportion of moths detected per morph for humans as the response variable and the proportion of moths detected per morph for each DNN independently as the explanatory variable. We fitted one linear model per DNN architecture. We considered that a DNN architecture could better predict human outcomes whenever significant coefficients were obtained and at higher r^2 values. Interactions between morph type and size were included by fitting independent linear models in three data subsets including different moth sizes: 0-20 pixels, 20-40 pixels and 40-60 pixels.

Results

Image elements used by DNNs to detect moths

Class activation maps (CAMs) showed that DNNs used features of both the wings and body to recognize moths. SqueezeNet produces the highest-resolution CAMs, so we show examples of CAMs generated by SqueezeNet for each of five morphs against two different backgrounds in Figure 1. SqueezeNet shows a strong bias towards the highest activations overlapping the anterior wing edges, with a slight bias toward the left anterior edge, regardless of whether the images were mirrored such that the left side became the right side (i.e., when the images were mirrored, the bias was toward the new left side that had previously been the right side.). The importance of diagonally-oriented wing edges is particularly evident in the bottom row of images in Figure 1, where high activations were produced in response to diagonally-oriented twigs viewed against a tree trunk. Although the actual moth was not detected in these images, the DNN still identified these images as having a moth in them, with probability scores ranging from 61-94%.

Comparison between DNN architectures

In the DNN experiment, AlexNet produced the highest false positive rate (0.24), while GoogLeNet produced the lowest (0.06, Table S1). Although DNN architectures ranked moths similarly, their responses were not identical when testing the different contrasts (Table 1). Opaque morphs were easier to detect than transparent ones for four out of six DNN architectures (AlexNet ($z = 4.26$, $p < 0.001$), GoogLeNet ($z = 14.6$, $p < 0.001$), ResNet ($z = 17.19$, $p < 0.001$) and SqueezeNet ($z = 7.04$, $p < 0.001$)). Moths with small windows were easier to detect than moths with large windows for four DNN architectures (GoogLeNet ($z = 4.34$, $p < 0.001$), ResNet ($z = 4.9$, $p < 0.001$), SqueezeNet ($z = 2.93$, $p < 0.005$) and vgg16 ($z = 2.89$, $p < 0.005$)). Only for GoogLeNet, morphs with transparent elements not touching any edge were more detected than morphs with transparent elements touching one moth edge (LW>BE: $z = 5.86$, $p < 0.001$). Breaking a single moth border instead of several increased moth detectability for all DNN architectures (BE>B3E: AlexNet: $z = 5.02$, $p < 0.001$, GoogLeNet: $z = 11.08$, $p < 0.001$, ResNet: $z = 6.32$, $p < 0.001$, SqueezeNet: $z = 8.57$, $p < 0.001$, vgg16: $z = 6.75$, $p < 0.001$, vgg19: $z = 3.49$, $p < 0.001$). Size increase improved detectability similarly for all DNN architectures (size: AlexNet: $z = 13.88$, $p < 0.001$, GoogLeNet: $z = 13.12$, $p < 0.001$, ResNet: $z = 10.55$, $p < 0.001$,

SqueezeNet: z = 14.38, p < 0.001, vgg16: z = 14.38, p < 0.001, vgg19: z = 16.16, p < 0.001). Opaque morphs were even more detected than transparent ones when they were large for four DNNs (AlexNet (z = 9.24, p < 0.001), SqueezeNet (z = 5.36, p < 0.001), vgg16 (z = 6.87, p < 0.001) and vgg19 (z = 10.31, p < 0.001)), but it was the opposite for ResNet (z = -3.45, p = 0.002). At larger sizes, moths with small windows gained more in detectability in comparison to moths with large windows for all DNN architectures (size:SW>LW: AlexNet: z = 4.95, p < 0.001, GoogLeNet: z = 4.07, p < 0.001, ResNet: z = 2.98, p < 0.005, SqueezeNet: z = 5.23, p < 0.001, vgg16: z = 5.95, p < 0.001, vgg19: z = 8.3, p < 0.001). Similarly, morphs with large transparent elements and complete borders gained more in detectability with size compared to morphs with one broken border for five out of the six tested DNN architectures (AlexNet: z = 2.6, p < 0.005, SqueezeNet: z = 3.23, p < 0.005, ResNet: z = 6.51, p < 0.001, vgg16: z = 7.09, p < 0.001, vgg19: z = 7.8, p < 0.001). However, detectability of moths with one instead of three broken edges increased with size only for ResNet (z = 5.33,/span>p <0.001), vgg16 (z = 5.09, p <0.001) and vgg19 (z = 8.36, p <0.001).

1

Table 1. Recapitulation of model results per DNN (first 6 columns) and humans (column 7). Whenever statistical results go in the direction of the relationship stated in the first column we wrote "yes. In case statistical results were significant but suggest the opposite direction we wrote "no". NS stands for non-significant differences (p-value threshold: 0.05). All results that agree with human results are shown in bold. * highlights only differences where none of the DNNs showed the same response as humans. Dots on the detectability ranking row mean undetermined order between morphs. Size was rescaled in all models. Tables summarized here can be found in ESM.

	AlexNet	GoogLeNet	ResNet	SqueezeNet	vgg16	vgg
O>all	**yes**	**yes**	**yes**	**yes**	NS	NS
SW>LW	NS	**yes**	**yes**	**yes**	**yes**	NS
LW>BE	NS	**yes**	NS	NS	NS	NS
BE>B3E	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**
size:O>all	yes	NS	**no**	yes	yes	yes
size:SW>LW	yes	yes	yes	yes	yes	yes
size:LW>BE	yes	**NS**	yes	yes	yes	yes
size:BE>B3E	NS	NS	yes	NS	yes	yes
Detectability ranking	O>SW.LW.BE>B3E	**O>SW>LW>BE>B3E**	O>SW>LW.BE>B3E	O>SW>LW.BE>B3E	O.SW>LW.BE>B3E	O.S

1

Comparison between DNNs' and humans' reactions

Humans had a similar false discovery rate to DNNs (Table S1): humans reported seeing a moth when facing 16% of photos without any moth (i.e., in between the 6% and 24% false discovery rate for DNNs). Humans timed-out responses when the picture had a moth where larger for B3E (67 responses) morph, followed by similar values for BE (42 responses) and LW (44 responses). The fewest timed-out responses were for C (6 responses) and SW (13 responses). Humans were better than DNNs at detecting moths (Fig. 2, human results are always above the diagonal). For both DNNs and humans, moths exhibiting transparent elements were detected in

most cases (score > 0.5, Figure 2). Morph detectability rankings (O>all, SW>LW, LW>BE and BE>B3E) were similar for humans and GoogLeNet, but not for the other individual DNN architectures (Tables 1, S2 and S3 and Figure 2).

Like DNNs, humans detected larger moths more easily (humans z =10.96, $p < 0.001$, Figure 2). However, the interaction between size and morph detectability in humans contrasted with the general pattern reported for DNNs. In humans, morphs that were more difficult to detect generally gained more in detectability with size increase than morphs that were easier to detect (Table S3), whereas the opposite was true of DNNs. I.e., morphs that were easier to detect for DNNs gained more in detectability with size increase than morphs that were more difficult to detect. For humans and opposite to AlexNet, SqueezeNet, vgg16 and vgg19, the gain in detection with size was larger for transparent moths than for opaque ones (size: O>all z = -8.62, $p < 0.001$). In contrast to all DNN results, at larger morph sizes, morphs with large windows gained more in detectability than morphs with small windows for humans (size: SW>LW: z = -6.6, $p < 0.001$). For humans and in contrast to all DNNs, the size-dependent gain in detection was greater for moths with multiple broken borders than for moths with one broken border (size: BE>B3E: z = -3.99, p <0.001). No difference in detectability gain was detected at larger morph sizes when comparing broken and unbroken borders (size:LW<BE: z = -1.25, p = 0.21), similar only to GoogLeNet. Finally, human participants performed slightly better by thend than at the beginning of the game (sequence: z = 7.27, p <0.001). Since DNNs do not learn during the testing phase, no comparison between humans and DNNs could be made here.

Predictability of human behaviour from DNN results varied according to morph size (Figure 2 and Table S4). Results of DNNs and humans were more similar for sizes between 0-20 pixels and especially 20-40 pixels than for sizes between 40-60 pixels. Additionally, DNN architectures differed in their predictability of human responses from low such as AlexNet to rather high such as GoogLeNet and vgg19 (Figure 2).

Discussion

Can DNNs be a reliable proxy for human perception?

In our experiments, we tested the detection of cryptic moths with or without transparent elements differing in size and position by six different DNN architectures and humans. Although the responses among observers were not identical, we can report some general trends shared by all observers. Humans and DNNs readily detected many of the prey items presented during the experiment, suggesting that for both of them our prey were not extremely cryptic. However, we can identify some key traits that reduced detectability for most of the observers: the presence of transparent elements and several broken moth borders. These traits correspond to detectability reducers that have already been documented elsewhere. Previous experiments with humans and birds have shown that transparent elements reduce prey detectability, probably because they enhance background matching (Arias et al., 2019, 2020) and/or resemble holes caused by decay or insect damage (Costello et al., 2020). Broken borders have also reduced detectability of prey by birds in other studies (Fig. S1, Arias et al., 2021; Cuthill et al., 2005), enhanced by inner background-matching elements (Fraser et al., 2007). Experiments using a computational vision model suggest that disruptive coloration affects edge-recognition algorithms enhancing detection of inner “edges” that hamper real edge detection and thus prey detectability (Stevens & Cuthill, 2006).

Despite these general trends, ranking was not identical among DNN architectures and humans. Human ranking of the different morphs was only shared with GoogLeNet, suggesting an unexpected variability of DNN architecture responses. These differences could be related to the salience of different traits for the different observers. Class activation maps suggest that diagonally-oriented wing edges were particularly important for moth detection, at least for SqueezeNet, and that in some cases, these edges on their own were sufficient to elicit a categorization of moth (Fig. 1). This suggests that DNNs may not have learned the general form of the moths, but rather some isolated diagnostic features, like diagonally-oriented edges. Salience differences between elements and DNN architectures may be related to the differences in DNN responses.

For our dataset, GoogLeNet was the most reliable (albeit still imperfect) proxy for detection by humans, but whether or not it would mirror human responses to a similar extent for other types of stimuli remains uncertain. Importantly, we found that the size of the moths had a large effect on the different morphs' relative detectability, such that GoogLeNet's responses no longer mirrored those of humans when the moths were large. For humans, the morphs that gained the most in detectability with size increase were those that were already more difficult to detect. This effect may have been driven, in part, by the fact that the morphs that were easier to detect were detected at a high rate even when small, so they could gain very little in detectability with size increaseAs a result, the detectability of all morphs seemed to reach a plateau at large sizes (Figure 2). By contrast, for most DNNs and most morph comparisons, size increase rendered more detectable all morphs but even more those that were already easier to detect (19/22 possible comparisons morph contrast x DNN architecture, Table 1). For DNNs, the relative detectability of the different morphs changed dramatically between the small and medium size classes, but very little between the medium and large size clsses, and did not appear to be approaching a plateau as was observed in the human data (Figure 2). This suggests that humans and DNNs may have differentially weighted and/or used different features or feature combinations whose saliencies are not perfectly correlated at different apparent viewing distances. Still, our dataset had fewer photos of moths of larger sizes (457 pictures below 20 pixels, 695 picture between 20 and 40 pixels and 155 pictures between 40 and 60 pixels) and this could have had an effect on DNNs and humans, so one could argue that training the DNNs on pictures with an uneven distribution of sizes could have led to this effect, although it is difficult to imagine the mechanism behind such an effect. It is also worth mentioning that these moth stimuli were novel to both humans and DNNs, and thus both types of observers exhibited learning – DNNs during their training and humans over the course of testing. Uneven size distribution could have had an effect on learning in both types of observers, but whether the effect would be the same for both humans and DNNs is unclear. For a novel set of stimuli, it would be difficult to make recommendations as to how large a target should be in order for GoogLeNet to mirror human responses, as the optimal size for a given stimulus class likely depends on a variety of unpredictable factors. In order to better understand the differential interactions between size and other target characteristics in humans versus DNNs, future studies could include items of varying visual complexity that are harder for humans to detect. It is interesting to note that Figure 2 suggests that when absolute detectability for humans and DNNs is similar (i.e., when the plotted points lie along the y=x line), the relative detectability of the different morphs is more likely to be similar (albeit still not identical). It would be interesting to test whether the same holds true for a completely different set of stimuli.

DNNs have made incredible strides in mimicking biological perception in the last decade, but they are still susceptible to error in ways that computer scientists are still discovering (Serre,

2019). For example, recent studies have shown that DNNs do not encode global object shape, but rather local object features, and perform poorly when asked to classify images composed of silhouettes without any internal object texture (Kubilius et al., 2016). Indeed, Baker et al. (2018) found VGG-19's performance to be abysmal when objects had clearly-defined shapes against a white background but incongruous internal textures. An interesting example was that of a shape of a vase with an internal texture taken from a photo of a gong. For a human, the vase shape is evident whereas the identity of the texture is unclear, whereas VGG-19 assigned the highest probability (22%) to "gong" and only a very low probability (1.8%) to "vase." This represents a key difference between human and DNN perception and may help explain why DNN and human responses did not perfectly mirror each other in the present study. Features that were diagnostic of "moth" likely differed between humans and DNNs in this experiment. For humans, the entire outline would have been important, whereas for DNNs, isolated features like a plain gray diagonal stripe seem to have been important.

Recent work has taken a different approach to using DNNs to study camouflage than the approach used here, and may represent a promising alternative approach that could potentially yield more biologically representative results. Rather than training DNNs on the presence versus absence of a target in an image, Fennell et al. (2019; 2021) trained DNNs on reaction times obtained from real humans in a detection task involving a limited range of phenotypes, from highly conspicuous to highly cryptic, and then asked DNNs to predict human reaction times to a large range of novel targets. To test whether DNNs' predictions for novel targets were accurate, humans were presented with a subset of these targets and their reaction times recorded. Specifically, humans were presented with targets that DNNs predicted would have very different reaction times (high vs. low (Fennell et al., 2021) or high vs. intermediate vs. low (Fennell et al., 2019)). Mean human reaction times between these coarse categories differed in the expected directions, but fine-scale validation tests on the relative detectability of similar-looking targets were not conducted; this remains an interesting avenue for future research. That said, such an approach would not have been a more efficient way to test specific hypotheses such as the ones tested in the present study because, unfortunately, it still requires significant input from human subjects.

Generalization to field trials with birds

While we found strong resemblance between at least one of the DNNs and human participants, their responses only partially matched those reported for birds (Arias et al., 2021). Whereas in birds, only the position but not the size of transparent windows seemed to affect prey detectability, in DNNs and humans, both the size and position of the windows strongly affected detectability.

One potential explanation for the lack of agreement between wild birds and humans and DNNs could be related to what was effectively being tested in our experiments. A real-life predation event requires several steps including prey detection, identification, handling and consumption (Endler, 1986). Yet human and DNN experiments only tested the detectability and identification of different items as potential prey (i.e., moths). By contrast, the field experiment only counted detections that were followed by the entire chain of steps, without the possibility to include events when prey was detected but not attacked. Therefore, conditions that can affect the motivation to attack, and thus reduce the number of attacked prey in comparison to detected prey, such as level of hunger (Savory et al., 1993; Savory & Lariviere, 2000) or neophobia (Marples & Kelly, 1999), moths resembled no local moth), were part of the field experiment with birds but not of the experiments with DNNs or humans.

DNN, human and bird experiments also differed in the heterogeneity of light conditions at which they were exposed to prey. The original bank of images that was later randomised and presented to humans and DNNs included many sets of pictures of the different morphs at constant background, morph size and light conditions per set. By contrast, avian predators were exposed to prey in the field where light conditions are highly diverse. Natural outdoor illumination varies not only across different sun elevations and weather conditions, but also across different viewing angles and microhabitats (Endler, 1993; Johnsen et al., 2006; Siegel et al., 1999; Tedore & Nilsson, 2019), and can be expected to vary quickly as a bird moves from one perch to another. Such variables may have introduced noise into the wild bird data and obscured differences in detectability that may have been observed under more controlled conditions. It would be interesting to run a laboratory-based follow-up experiment in which birds are trained to peck moths in images on touch screens, as well as a field-based experiment in which humans are asked to search for artificial moths in a real forest. Perhaps if the conditions experienced by humans, birds, and DNNs were more similar, their results would also have been more similar. That said, the conditions experienced by birds and humans in the forest could never be made entirely equivalent due to the different ways humans and birds move through the environment.

The spatial configuration of elements seen by birds may have also differed from that seen in the photos viewed by humans and DNNs. Birds may search for moths from a different viewing angle than the moths were photographed from. All photographs were taken from a viewing angle perpendicular to the surface of the moth wings. By contrast, birds may search while flying or walking on the trunk or the ground. Different viewing angles can be expected to impact not only the apparent configuration of opaque elements, but also the specular reflection (i.e., glare) by transparent elements. In the first set of images in Figure 1, some specular reflection can be seen from the transparent windows of the BE morph. Real butterflies and moths with transparent elements show comparatively less specular reflection; thus, it would be interesting to run a similar set of experiments with artificial moths containing empty windows rather than transparent windows made from synthetic materials.

However, it is possible that differences in the results are not related to the experimental set-up, but rather to the intrinsic functioning of the different observer "visual" systems, such as differences in the colour and achromatic contrasts seen by the different observers. DNNs and humans were fed standard RGB photos from consumer cameras designed to produce reasonably realistic color and achromatic contrasts to humans. Birds, on the other hand, possess a rather different visual system with four cones that are differently-tuned and more spectrally separated than the three cones of humans (Serre, 2019). Although calculations predicted low color and achromatic contrasts between moths and tree trunks for birds (Arias et al., 2021), it is unlikely that the contrasts in our images perfectly replicated those seen by birds in the field. Indeed, the coor and especially achromatic channels of birds are known to be much noisier than those of humans (Olsson et al., 2018). The noisier the visual system, the more likely it is that the wing edges bordering a transparent window will be obscured by noise. It is possible that, when viewed from a distance, the edges of SW and LW were similarly obscured for birds, but not for humans or DNNs. It would be useful to test whether programming and training DNNs from scratch that accept multispectral images with four or even five channels (in order to include the achromatic double cone channel), taken by a camera mimicking bird vision (see, e.g., camera with custom filters described in Tedore and Nilsson (2019, 2021)), with noise mimicking that of avian cones introduced into the images, would yield more similar results to those obtained from wild birds.

Finally, it is worth mentioning that the sample size of the field experiment with birds (1733 moths, of which 618 were attacked) was over an order of magnitude lower than the sample

sizes obtained in the present study from humans (N=31130 picture views) and DNNs (N=23400 picture views). Although 618 attacks is a high number for a field study, it is possible that with an even larger sample size, some small effect of transparent element size would have been detectable.

Conclusions and future directions

DNNs and humans showed similar but not identical reactions towards the same stimuli in our experiments. Strong resemblance was found between GoogLeNet and human morph ranking. However, the apparent distance (size) of the prey affected the strength of the effects of the size and position of transparent elements, with small prey size generally strengthening these effects in humans and weakening them in DNNs. With the six DNNs used in our study, we were unable to simulate humans' sensitivity to object size. This suggests that prey size/viewing distance can interact with camouflage type in opposing directions in humans and DNNs, and warrants a fuller and more targeted investigation of size interactions with a broader range of stimuli.

Neither human nor DNN responses closely matched those of wild birds in the field. There were several potentially confounding factors between field and lab conditions, however. Future work should better control for viewing angle, lighting, and motivation in bird experiments by using laboratory-trained birds. Future work should also compare the DNN results obtained in this study to those obtained with a DNN that accepts images with 4+ channels taken with a multispectral camera mimicking avian vision such as those in Tedore and Nilsson (2019, 2021).

Data availability

The datasets generated during the current study are available from the corresponding author on reasonable request.

Conflict of Interest

We have no conflict of interest to declare.

Acknowledgements

This work was funded by Clearwing ANR project (ANR-16-CE02-0012) and HFSP project on transparency (RGP0014/2016). With the support of LabEx CeMEB, an ANR "Investissements d'avenir" program (ANR-10-LABX-04-01).

References

Arias, M., Elias, M., Andraud, C., Berthier, S., & Gomez, D. (2020). Transparency improves concealment in cryptically coloured moths. *Journal of Evolutionary Biology*, *33*(2), 247–252.

Arias, M., Leroy, L., Madec, C., Matos, L., Tedore, C., Elias, M., & Gomez, D. (2021). Partial wing transparency works better when disrupting wing edges: Evidence from a field experiment. *Journal of Evolutionary Biology*, *34*(11), 1840–1846.

Arias, M., Mappes, J., Desbois, C., Gordon, S., McClure, M., Elias, M., Nokelainen, O., & Gomez, D. (2019). Transparency reduces predator detection in mimetic clearwing butterflies. *Functional Ecology*, *33*(6), 1110–1119.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), e1006613.

Beatty, C. D., Bain, R. S., & Sherratt, T. N. (2005). The evolution of aggregation in profitable and unprofitable prey. *Animal Behaviour*, *70*(1), 199–208.

Bond, A. B., & Kamil, A. C. (1998). Apostatic selection by blue jays produces balanced polymorphism in virtual prey. *Nature*, *395*(6702), 594–596.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput Biol*, *10*(12), e1003963.

Charpentier, M., Harté, M., Poirotte, C., de Bellefon, J. M., Laubi, B., Kappeler, P., & Renoult, J. (2020). Same father, same face: Deep learning reveals selection for signaling kinship in a wild primate. *Science Advances*, *6*(22), eaba3274.

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, *10*(10), 1632–1644.

Costello, L. M., Scott-Samuel, N. E., Kjernsmo, K., & Cuthill, I. C. (2020). False holes as camouflage. Proceedings. Biological Sciences, 287(1922), 20200126.

Cox, D. R. (1972). Models and life-tables regression. *JR Stat. Soc. Ser. B*, *34*, 187–220.

Cuthill, I. C., Stevens, M., Sheppard, J., Maddocks, T., Párraga, C. A., & Troscianko, T. S. (2005). Disruptive coloration and background pattern matching. *Nature*, *434*(7029), 72.

Cuthill, J. F. H., Guttenberg, N., Ledger, S., Crowther, R., & Huertas, B. (2019). Deep learning on butterfly phenotypes tests evolution's oldest mathematical model. *Science Advances*, *5*(8), eaaw4967.

Endler, J. A. (1986). Defense against predation. In *Predator-prey Relationships, Perspectives and Approaches from the Study of Lower Vertebrates. In M. E. Feder and G. E. Lauder [eds.]* (pp. 109–134). University of Chicago Press.

Endler, J. A. (1993). The Color of Light in Forests and Its Implications. *Ecological Monographs*, *63*(1), 2–27. https://doi.org/10.2307/2937121

Finkbeiner, S. D., Briscoe, A. D., & Reed, R. D. (2012). The benefit of being a social butterfly: Communal roosting deters predation. *Proceedings of the Royal Society B-Biological Sciences*, *279*(1739), 2769–2776. https://doi.org/10.1098/rspb.2012.0203

Fraser, S., Callahan, A., Klassen, D., & Sherratt, T. N. (2007). Empirical tests of the role of disruptive coloration in reducing detectability. *Proceedings of the Royal Society of London B: Biological Sciences*, *274*(1615), 1325–1331.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Johnsen, S., Kelber, A., Warrant, E., Sweeney, A. M., Widder, E. A., Lee Jr, R. L., & Hernández-Andrés, J. (2006). Crepuscular and nocturnal illumination and its effects on color perception by the nocturnal hawkmoth Deilephila elpenor. *Journal of Experimental Biology*, *209*(5), 789–800.

Kazemi, B., Gamberale-Stille, G., Tullberg, B. S., & Leimar, O. (2014). Stimulus salience as an explanation for imperfect mimicry. *Current Biology*, *24*(9), 965–969.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, *12*(4), e1004896.

Marples, N. M., & Kelly, D. J. (1999). Neophobia and Dietary Conservatism:Two Distinct Processes? *Evolutionary Ecology*, *13*(7–8), 641–653. https://doi.org/10.1023/A:1011077731153

Mély, D. A., & Serre, T. (2017). Towards a theory of computation in the visual cortex. In *Computational and cognitive neuroscience of vision* (pp. 59–84). Springer.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, *115*(25), E5716–E5725.

Olsson, P., Lind, O., & Kelber, A. (2018). Chromatic and achromatic vision: Parameter choice and limitations for reliable model predictions. *Behavioral Ecology*, *29*(2), 273–282.

R Foundation for. (2014). *R: A language and environment for statistical computing.* [Computer software].

Renoult, J. P., Guyl, B., Mendelson, T. C., Percher, A., Dorignac, J., Geniet, F., & François, M. (2019). Modelling the Perception of Colour Patterns in Vertebrates with HMAX. *bioRxiv*, 552307. https://doi.org/10.1101/552307

Savory, C., & Lariviere, J.-M. (2000). Effects of qualitative and quantitative food restriction treatments on feeding motivational state and general activity level of growing broiler breeders. *Applied Animal Behaviour Science, 69*(2), 135–147.

Savory, C., Maros, K., & Rutter, S. (1993). Assessment of hunger in growing broiler breeders in relation to a commercial restricted feeding programme. *Animal Welfare, 2*(2), 131–152.

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science, 5*, 399–426.

Sherratt, T. N., Whissell, E., Webster, R., & Kikuchi, D. W. (2015). Hierarchical overshadowing of stimuli and its role in mimicry evolution. *Animal Behaviour, 108*, 73–79. https://doi.org/10.1016/j.anbehav.2015.07.011

Siegel, D. A., Westberry, T. K., & Ohlmann, J. C. (1999). Cloud color and ocean radiant heating. *Journal of Climate, 12*(4), 1101–1116.

Stevens, M., & Cuthill, I. C. (2006). Disruptive coloration, crypsis and edge detection in early visual processing. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1598), 2141. https://doi.org/10.1098/rspb.2006.3556

Talas, L., Fennell, J. G., Kjernsmo, K., Cuthill, I. C., Scott-Samuel, N. E., & Baddeley, R. J. (2020). CamoGAN: Evolving optimum camouflage with Generative Adversarial Networks. Methods in Ecology and Evolution, 11(2), 240–247.

Tedore, C., & Nilsson, D.-E. (2019). Avian UV vision enhances leaf surface contrasts in forest environments. *Nature Communications*, *10*(1), 1–12.

Tedore, C., & Nilsson, D.-E. (2021). Ultraviolet vision aids the detection of nutrient-dense non-signaling plant foods. *Vision Research, 183*, 16–29.

The MathWorks In. (2019). *MATLAB* (Version 9.7.0.1296695 (r2019b)) [Computer software].

Therneau, T. M. (2020). *Mixed Effects Cox Models [R package coxme version 2.2-16]*.

Tosh, C. R., & Ruxton, G. D. (2010). *Modelling perception with artificial neural networks*. Cambridge University Press.

Troscianko, J., Lown, A. E., Hughes, A. E., & Stevens, M. (2013). Defeating crypsis: Detection and learning of camouflage strategies. *PloS One*, *8*(9), e73733.

Tucker, G., & Allen, J. (1988). Apostatic selection by humans searching for computer-generated images on a colour monitor. *Heredity*, *60*(3), 329–334.

Wham, D. C., Ezray, B., & Hines, H. M. (2019). Measuring Perceptual Distance of Organismal Color Pattern using the Features of Deep Neural Networks. *bioRxiv*, 736306. https://doi.org/10.1101/736306

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Ying, X. (2019). *An overview of overfitting and its solutions*. *1168*(2), 022022.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). *The unreasonable effectiveness of deep features as a perceptual metric*. 586–595.

1

Caption figures

Figure 1. Example images of each of the five morphs against two different backgrounds (a. and b.) at different viewing distances, together with their corresponding class activation maps (CAMs) produced by SqueezeNet. The CAMs are heatmaps with “hot”, or red, regions corresponding to parts of the image with the highest activations (centre of the clouds displayed in the second and fourth rows of pictures). Note that in (a), the DNN has correctly identified the locations of the moths, whereas in (b), the DNN consistently misidentifies two crossed diagonally-oriented twigs as a moth.

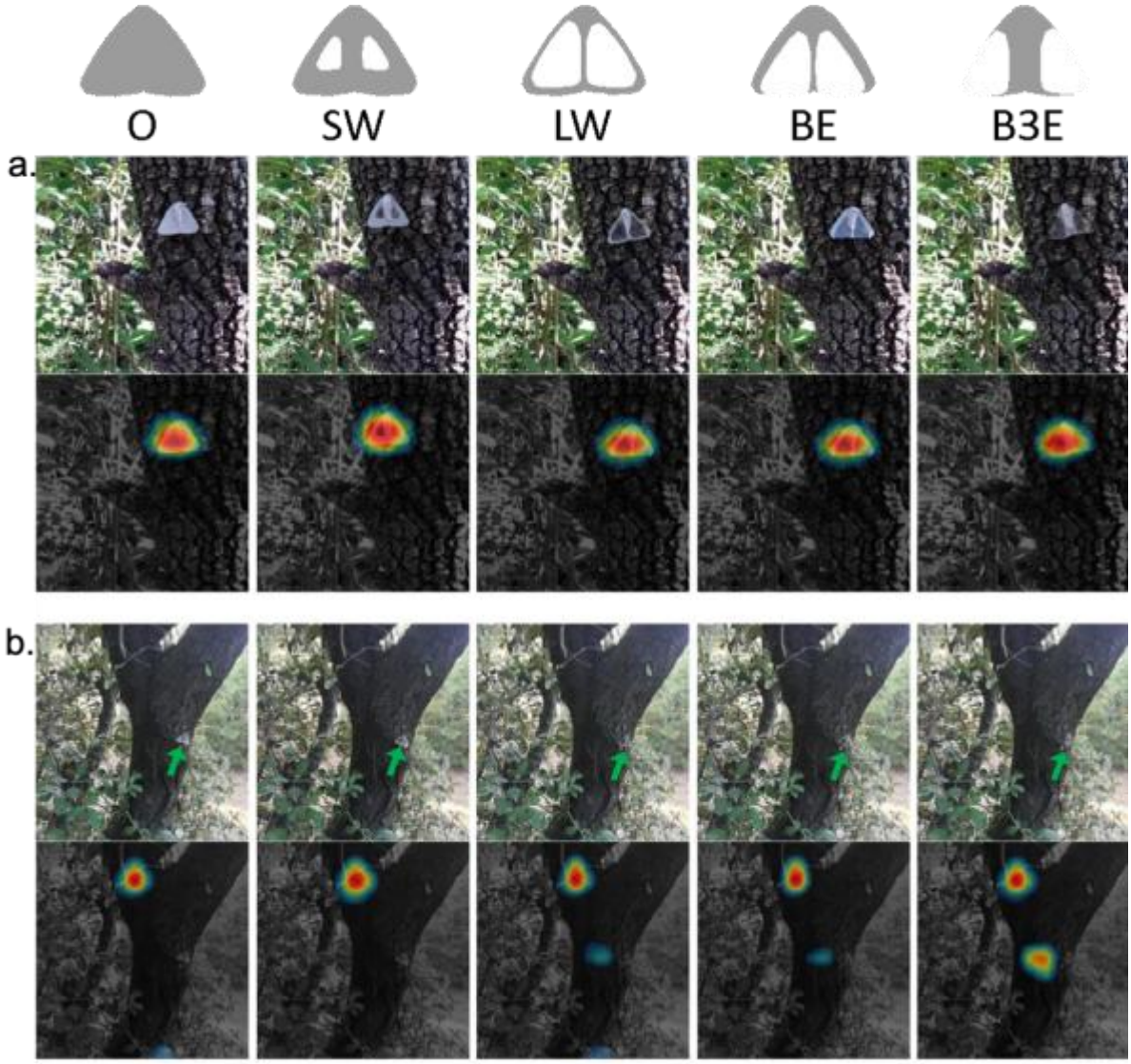

Figure 2. Mean detection of the different morphs by AlexNet (a, b, c), GoogLeNet (d, e, f), ResNet (g, h, i), SqueezeNet (j, k, l), vgg16 (m, n, o) or vgg19 (p, q, r) (x- axis) and by humans (y-axis) at different moth sizes: below 20 pixels (a, d, g, j, m, p), between 20 and 40 pixels (b, e, h, k, n, q) and from 40 to 60 pixels (c, f, i, l, o, r). Diagonal represents what would be identical responses for humans and DNNs.

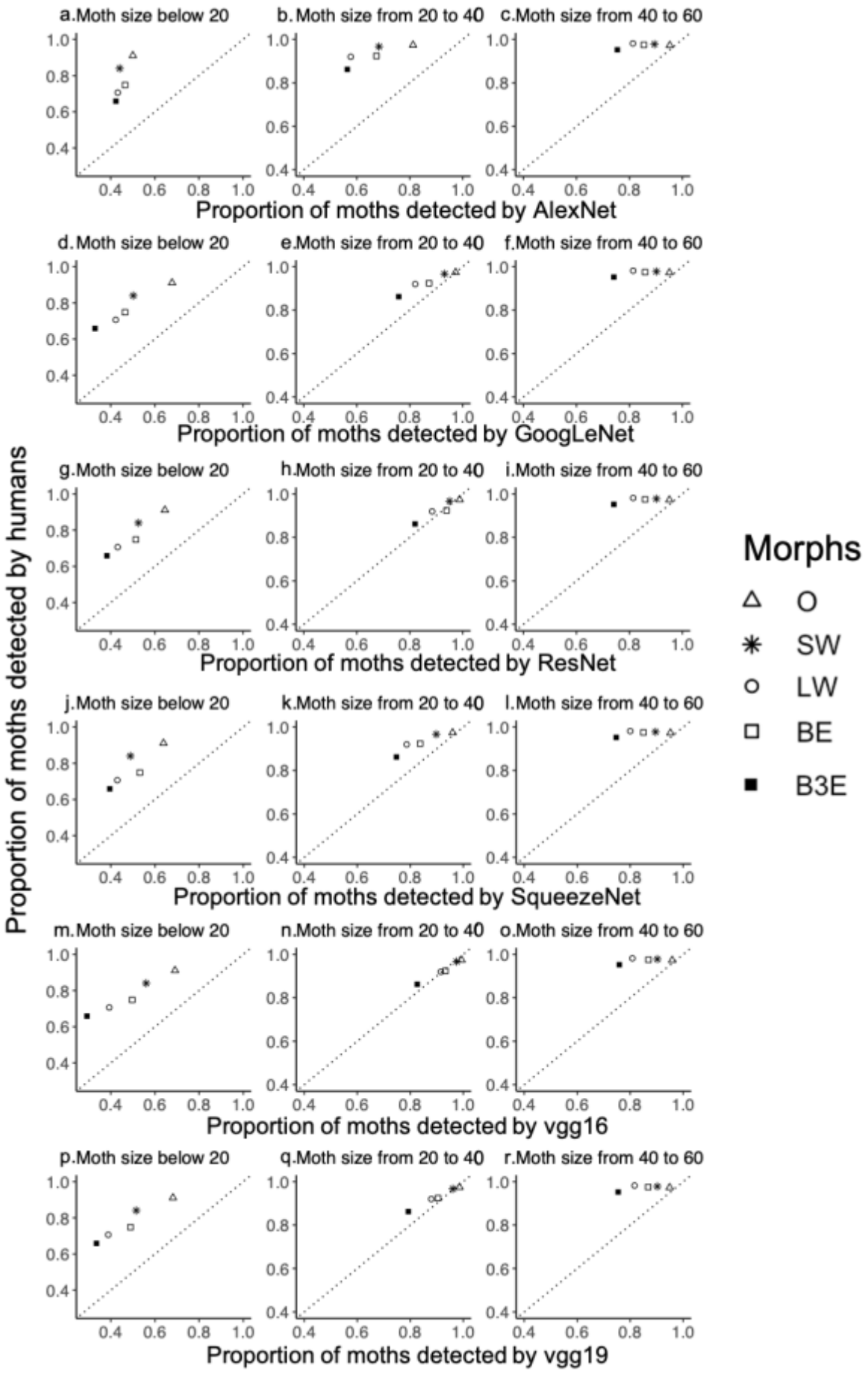
a.Moth size below 20
b.Moth size from 20 to 40
c.Moth size from 40 to 60
Proportion of moths detected by AlexNet
d.Moth size below 20
e.Moth size from 20 to 40
f.Moth size from 40 to 60
Proportion of moths detected by GoogLeNet
g.Moth size below 20
h.Moth size from 20 to 40
i.Moth size from 40 to 60
Proportion of moths detected by ResNet
j.Moth size below 20
k.Moth size from 20 to 40
l.Moth size from 40 to 60
Proportion of moths detected by SqueezeNet
m.Moth size below 20
n.Moth size from 20 to 40
o.Moth size from 40 to 60
Proportion of moths detected by vgg16
p.Moth size below 20
q.Moth size from 20 to 40
r.Moth size from 40 to 60
Proportion of moths detected by vgg19
Proportion of moths detected by humans
0.4
0.6
0.8
1.0
Morphs
O
SW
LW
BE
B3E

Figure S1. Morph survival when exposed to wild avian predators under field conditions. On such experiment artificial moths had wings made with paper and plastic wings and a fully edible body made with flour, water, lard and edible ink. Moths sporting the five different morphs also used in the current study, were placed on tree trunks and their 'survival' was monitored every day for four days. Bird attacks were detected by the marks left on the body. Figure and data from Arias et al (2021).

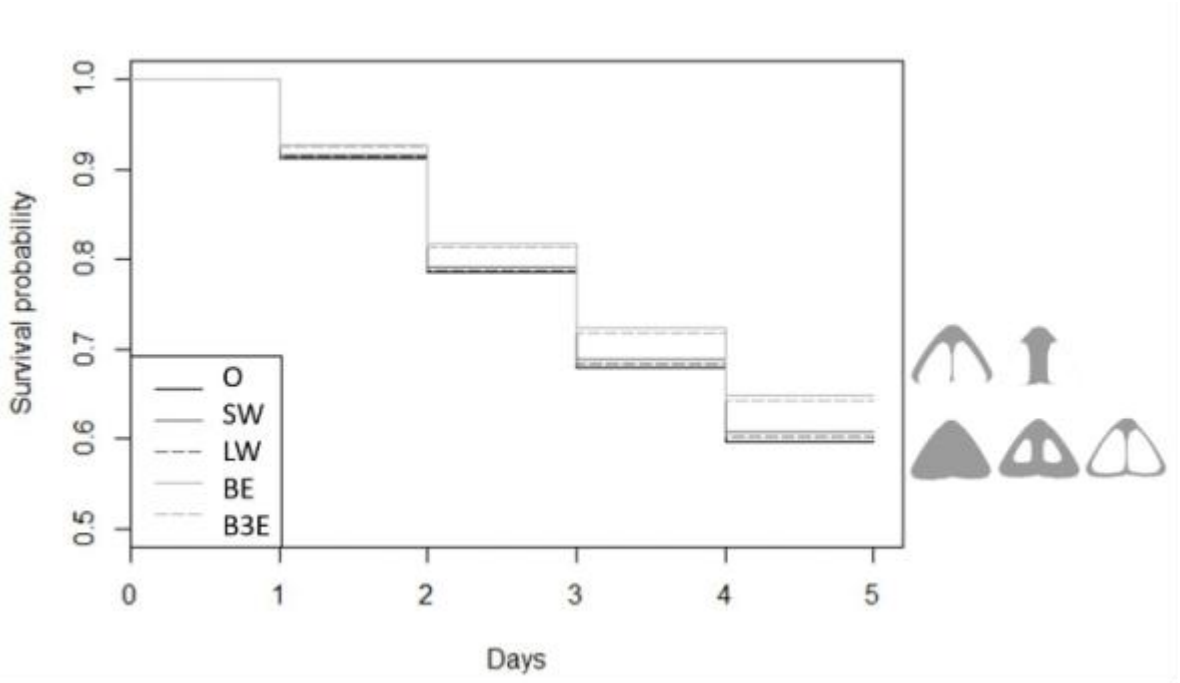

Table S1. False positive rate for different DNNs and humans

Type	Architecture	False positive rate
DNN	SqueezeNet	0.11
	ResNet	0.09
	AlexNet	0.24
	vgg16	0.09
	vgg19	0.12
	GoogLeNet	0.06
Humans		0.16

Table S2: GLMM results for each DNN. Size was rescaled in all models

	AlexNet			GoogLeNet			ResNet		
	Estim±SE	*z*	*p*	Estim±SE	*z*	*p*	Estim±SE	*z*	*p*
Intercept	-2.05±0.28	-7.29	<0.001	-2.26±0.49	-4.61	<0.001	-0.85±0.56	-1.52	0.129
morphs.O>all	0.11±0.03	4.26	<0.001	0.61±0.04	14.6	<0.001	0.48±0.03	17.19	<0.001
morphs.SW>LW	0.13±0.08	1.62	0.105	0.66±0.15	4.34	<0.001	0.55±0.11	4.9	<0.001
morphs.LW> BE	-0.09±0.09	-0.98	0.329	0.85±0.15	5.86	<0.001	0.19±0.12	1.54	0.124
morphs.BE>B3E	0.4±0.08	5.02	<0.001	1.37±0.12	11.08	<0.001	0.73±0.12	6.32	<0.001
size	10.86±0.78	13.88	<0.001	20.81±1.59	13.12	<0.001	19.78±1.87	10.55	<0.001
size:morphs.O>all	0.99±0.11	9.24	<0.001	-0.16±0.22	-0.76	0.448	-0.39±0.11	-3.45	<0.001
size:morphs.SW>LW	1.38±0.28	4.95	<0.001	3.01±0.74	4.07	<0.001	1.49±0.5	2.98	<0.005
size:morphs.LW>BE	0.77±0.3	2.6	<0.005	0.75±0.63	1.2	0.231	1.69±0.52	3.23	<0.005
size:morphs.BE>B3E	0.36±0.25	1.46	0.144	0.22±0.47	0.47	0.640	2.49±0.47	5.33	<0.001

	SqueezeNet			vgg16			vgg19		
	Estim±SE	*z*	*p*	Estim±SE	*z*	*p*	Estim±SE	*z*	*p*
Intercept	-2.46±0.42	-5.85	<0.001	-3.98±0.65	-6.12	<0.001	-2.8±0.51	-5.47	<0.001
morphs.O>all	0.29±0.04	7.04	<0.001	0.12±0.07	1.73	0.083	-0.05±0.05	-1.01	0.314
morphs.SW>LW	0.38±0.13	2.93	<0.005	0.55±0.19	2.89	<0.005	0.08±0.15	0.54	0.588
morphs.LW> BE	-0.15±0.13	-1.16	0.245	0.1±0.19	0.52	0.605	-0.1±0.16	-0.66	0.511
morphs.BE>B3E	0.93±0.11	8.57	<0.001	1.18±0.17	6.75	<0.001	0.46±0.13	3.49	<0.001
size	19.27±1.34	14.38	<0.001	39.22±2.73	14.38	<0.001	30.47±1.89	16.16	<0.001
size:morphs.O>all	1.16±0.22	5.36	<0.001	4.12±0.6	6.87	<0.001	4.85±0.47	10.31	<0.001
size:morphs.SW>LW	2.98±0.57	5.23	<0.001	7.31±1.23	5.95	<0.001	7.33±0.88	8.3	<0.001
size:morphs.LW> BE	3.36±0.52	6.51	<0.001	7.8±1.1	7.09	<0.001	6.42±0.82	7.8	<0.001
size:morphs.BE>B3E	0.18±0.39	0.45	0.651	4.39±0.86	5.09	<0.001	5.29±0.63	8.36	<0.001

Table S3: Cox Regression test with mixed effects on moth 'survival' (moth detection and time spent by participants at finding them). Random effects included picture set and participant code. Size was rescaled in all models

Humans

	Estim±SE	z	P	
morph.O>all	0.186±0.009	19.92	<0.001	***
morph.SW>LW	0.611±0.036	16.82	<0.001	***
morph.LW> BE	0.502±0.044	11.49	<0.001	***
morph.BE>B3E	0.599±0.039	15.25	<0.001	***
Size (i.e., apparent distance)	3.342±0.305	10.96	<0.001	***
Sequence (only humans)	0.002±0.0003	7.27	<0.001	***
Height (only humans)	0.002±0.001	1.38	0.170	
morph.O>all:size	-0.228±0.0264	-8.62	<0.001	***
morph.SW>LW:size	-0.667±0.101	-6.6	<0.001	***
morph.LW> BE: size	-0.151±0.121	-1.25	0.210	
morph.BE>B3E:size	-0.423±0.107	-3.93	<0.001	***

Table S4. Linear models showing the relationship between human reactions and each DNN architecture outcomes for three different moth size intervals. In bold are significant coefficient values. R^2 values correspond to adjusted r^2.

	0-20 size units				20-40 size units				40-60 size units			
	coef±S.Error	*t*	*p*	r^2	coef±S.Error	*t*	*p*	r^2	coef±S.Error	*t*	*p*	r^2
AlexNet	2.64±1.13	2.33	0.1	0.53	0.37±0.15	2.56	0.08	0.58	0.08±0.07	1.13	0.34	0.06
GoogLeNet	**0.76±0.13**	5.81	0.01	0.89	**0.51±0.08**	6.71	0.01	0.92	0.09±0.07	1.3	0.28	0.15
ResNet	**0.97±0.16**	5.94	0.01	0.89	**0.65±0.12**	5.23	0.01	0.87	0.08±0.07	1.17	0.33	0.08
SqueezeNet	**0.95±0.29**	3.24	0.05	0.7	**0.49±0.11**	4.54	0.02	0.83	0.07±0.07	1.05	0.37	0.02
vgg16	**0.65±0.07**	8.74	<0.001	0.95	**0.68±0.05**	14.56	<0.001	0.98	0.08±0.07	1.05	0.37	0.02
vgg19	**0.74±0.12**	6.41	0.01	0.9	**0.59±0.04**	15.88	<0.001	0.98	0.09±0.07	1.23	0.31	0.11