Natural warning signals unexpectedly shape human metamemory ratings but not image recognition success

Federico De Filippi[1,*], **Olivier Penacchio**[2,3], **Akira R. O'Connor**[1], **Julie M. Harris**[1]

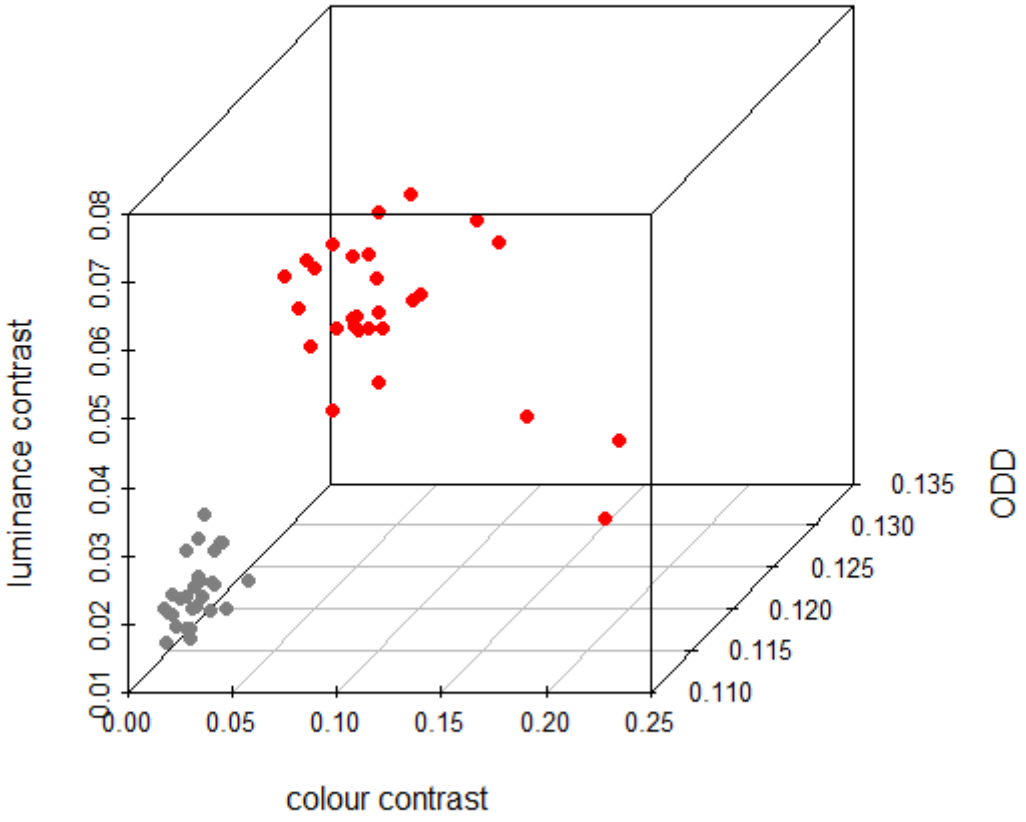
1. School of Psychology & Neuroscience, University of St Andrews, St Andrews KY16 9JP, United Kingdom
2. Bridging Research in AI and Neuroscience, Computer Vision Center, Cerdanyola del Vallès, 08193 Barcelona, Spain
3. Computer Science Department, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, 08193 Barcelona, Spain

* Corresponding author (fdf1@st-andrews.ac.uk)

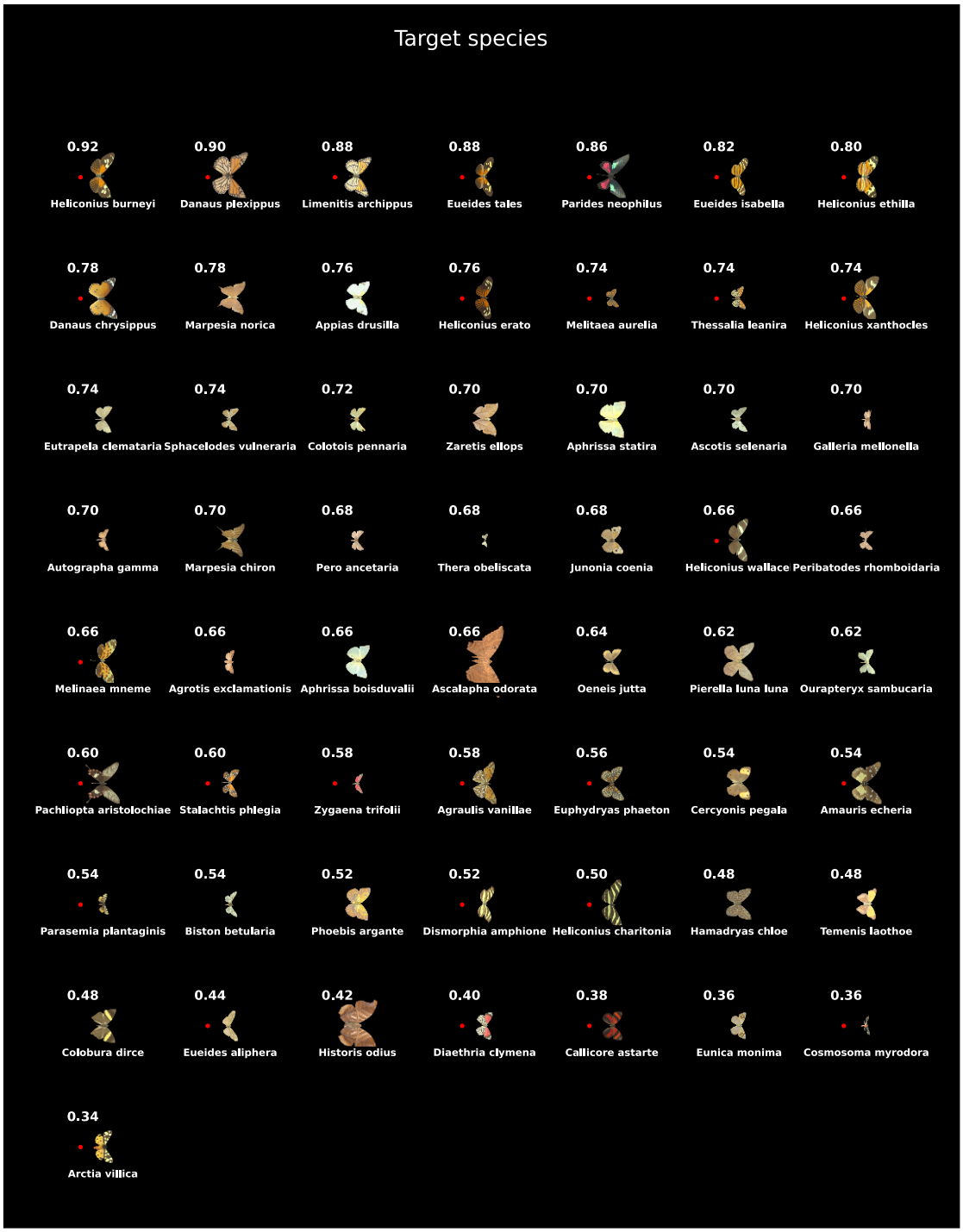
Supplementary information

Selection of species for memory experiment

All images implemented in the study were selected from the publicly available St Andrews database of hyperspectral images of Lepidoptera (https://arts.st-andrews.ac.uk/lepidoptera/documentation.html). For details on image acquisition, see Penacchio et al.[1] contains images of 125 species from 12 lepidopteran families: 96 aposematic (AP) and 29 non-aposematic (non-AP), sampled from museum collections located in British museums. For each image, the database also provides three metrics (luminance contrast, colour contrast, Orientation Distribution Deviation (ODD)) that characterise modelled activity in an avian visual system, according to the computational framework proposed by Penacchio et al [1]. As reported in the paper, these metrics can be used to classify butterflies as AP vs. non-AP with high accuracy. We fitted a logistic regression model to the database of neural metrics using the *R* [2] function *glm*, with luminance contrast, colour contrast, and ODD as predictors, and class (AP vs. non-AP) as the outcome variable. The logistic regression model was used to predict the overall neural signature for each image using the *R* [2] function *predict*. In other words, we used the predictive log odds that a given specimen is aposematic based on its visual statistics as a metric to select a set of stimuli that we presented to humans in the recognition memory test. When visualised in a three-dimensional pattern space (shown in *Supplementary Figure S1*) the selected stimuli occupy different regions of the space. Species that were used as memorisation targets are shown in *Supplementary Figure S2*, and lure species are shown in *Supplementary Figure S3*.



Supplementary Figure S1. Three-dimensional scatterplot of luminance contrast, colour contrast, and Orientation Distribution Departure for the Lepidoptera species (29 AP, 29 non-AP) from the St Andrews Hyperspectral Lepidoptera database [7] selected for the human memory experiment. Each point is a representative image of one species, selected based on a weighted sum of luminance contrast, colour contrast, and Orientation Distribution Deviation (ODD).



Supplementary Figure S2. Target species used in the human memory experiment, sorted by the hit rate (shown at the top-left of each species). The red dot marks aposematic species. The scientific name for each target is shown below the corresponding image.

Supplementary Figure S3. Lure species used in the human memory experiment, sorted by the false alarm rate (shown at the top-left of each species). The red dot marks aposematic species. The scientific name for each lure is shown below the corresponding image.

Conversion of hyperspectral images to sRGB colour space

The hyperspectral images from the St Andrews Hyperspectral Lepidoptera Database contain continuous information about the spectral reflectance of each pixel for several wavelengths, and as such, they required processing to be displayed as images in the conventional colour space. As proposed by Foster & Amano [3], reflectance data was converted to reflected radiance data by multiplying each pixel by the corresponding spectrum of daylight (correlated colour temperature (CCT) of 6500 K). Radiance data was subsequently converted to the RGB colour space using the CIE 1931 colour matching functions [4]. Specimens were segmented from the original hyperspectral scans and superimposed on a background with R, G, B values of 0 ('black'). This resulted in a database of pictures of single specimens of various sizes. For presentation, the converted mages were resized to a resolution of 256x256 using bilinear interpolation and a constant scaling factor of 256/*k*, where *k* is the size of the largest image in the database. Given the evidence of common frequency components in Lepidopteran wing patterns [1], this was done to preserve the natural size differences between specimens. The size of images in the online PsychoPy [5] experiment was not controlled, but stimuli were set to occupy half of participants' display and to be invariant to changes in aspect ratio of the window or display. To assess how differences in specimen sizes influenced the experimental data, we regressed the specimen sizes against the average metamemory ratings and hit rates. Size had a significant effect on metamemory ratings ($\beta = 1.36 \times 10^{-5}$, $t = 3.56$, $SE = 3.83 \times 10^{-6}$, $R^2 = 0.19$, $p < .001$) but not on hit rates ($\beta = 4.40 \times 10^{-6}$, $t = 0.85$, $SE = 5.21 \times 10^{-6}$, $R^2 = 0.01$, $p = .402$).

Statistical inference

To identify the most appropriate mixed-effects structures to model the experimental data, we fitted a series of mixed-effects models and compared them using likelihood-ratio tests. The tables below report the model selection results. *Supplementary Table S4* reports model comparisons for metamemory ratings, and *Supplementary Table S5* reports model comparisons for recognition accuracy.

Supplementary Table S4. Likelihood-ratio model comparisons for metamemory rating models. The model highlighted in grey was reported in the paper.

Model	Effects		AIC	BIC	Log Likelihood	χ^2	df	p-value
	Fixed	Random observer ID (experimental setting)						
Model 1 (null)		Intercept	-741.62	-723.76	373.81			
Model 2	+ Lepidoptera class	Intercept	-1,220.05	-1,196.23	614.02	480.42	1	< .001
Model 3	+ Lepidoptera class	Intercept + Lepidoptera class	-1,423.05	-1,387.32	717.52	207.00	2	< .001

Fixed effects				
	Estimate	**SE**	**t-value**	**p-value**
Intercept	0.451	0.224	20.118	< .001
Lepidoptera class	0.162	0.019	8.745	< .001

Random effects		
	Variance	**SD**
Intercept	0.452	0.155
Lepidoptera class	0.150	0.122

Model fit: R^2 (marginal): 0.108; R^2 (conditional): 0.475

Supplementary Table S5. Likelihood-ratio model comparisons for recognition accuracy models. The model highlighted in grey was reported in the paper.

Model	Effects		AIC	BIC	Log Likelihood	χ^2	df	p-value
	Fixed	Random observer ID (experimental setting)						
Model 1 (null)		Intercept	3584.70	3596.60	-1790.40			
Model 2	+ Lepidoptera class	Intercept	3586.10	3603.90	-1790.00	0.658	1	0.417
Model 3	+ Lepidoptera class	Intercept + Lepidoptera class	3565.00	3594.80	-1777.50	25.044	2	< .001

Fixed effects

	Estimate	SE	z-value	p-value
Intercept	0.602	0.123	4.883	< .001
Lepidoptera class	0.057	0.127	0.453	0.651

Random effects

	Variance	SD
Intercept	0.585	0.765
Lepidoptera class	0.456	0.675

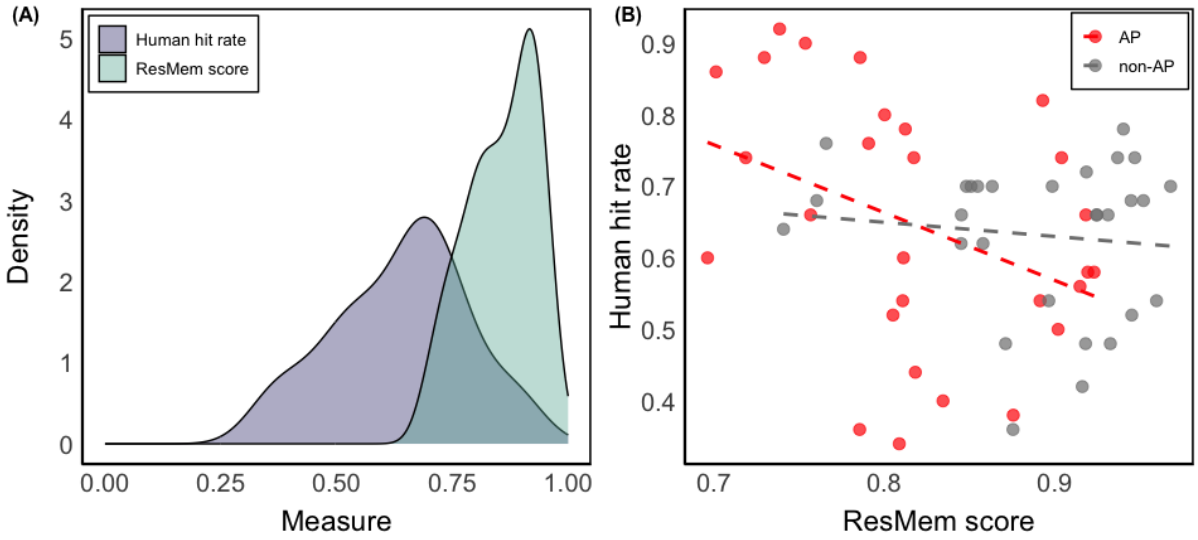
Model fit: R^2 (marginal): 0.000; R^2 (conditional): 0.108

ResMem analysis

Current memorability studies commonly test whether human hit rates can be predicted by computer-vision models such as ResMem [6]. These models assume that the high consistency observed across people (as we note in the main body of the paper) makes memory partly predictable from image features. Given our findings, we assessed whether this was the case for our Lepidoptera images.

The ResMem model [6] is a residual neural network pre-trained to estimate the memorability of any image. Its architecture combines a branch using AlexNet [7], an 8-layer deep neural network which is known to successfully predict object categories, and ResNet-152 [8], a 152-layer deep neural work which is known to successfully extract semantic information from images. The model is trained on datasets of complex images labelled by human memory performance, LaMem [9] and MemCat [10]. ResMem has been shown to correlate to human memory data (e.g., Davis & Bainbridge [11]). For our analysis, we initialised a pre-trained instance of ResMem from the python package *resmem* [6]. To pre-process the images for model inference, we used the *transformer()* function included with the package, which resizes the input to a 256x256 image, takes a 227x277 central crop, and converts the image to a tensor before passing it as input to ResMem. The model accepts any image as input and returns a 0–1 "memorability score", loosely interpretable as a predicted human hit rate. We exposed the model to our Lepidoptera images and compared its scores with empirical hit rates.

ResMem largely failed to match the distribution of human hit rates, producing a narrower range with an apparent bias toward high values (see *Supplementary Figure S6*). Across all species, ResMem scores were not significantly correlated with hit rates (Spearman's $\rho = -0.21$, $p = 0.38$). For AP species the correlation was significantly negative ($\rho = -0.41$, $p = .03$), and for non-AP species it was not significant ($\rho = 0.04$, $p = 0.84$). Thus, although human hit rates were highly predictable across observers, they were not well predicted by ResMem, which produced trends opposite to those observed in human data.



Supplementary Figure S6. Correspondence between artificial memorability scores obtained from ResMem and empirical hit rates for each Lepidoptera target species. (**A**) Distributions of human hit rates and ResMem scores across all species. (**B**) Scatterplot and best linear fits of the correlation between ResMem scores and human hit rates, grouped by Lepidoptera class.

References

1. Penacchio, O. *et al.* A computational neuroscience framework for quantifying warning signals. *Methods in Ecology and Evolution* **15**, 103–116 (2024).
2. R Foundation for Statistical Computing, R. C. T. R: A language and environment for statistical computing. (2021).
3. Foster, D. H. & Amano, K. Hyperspectral imaging in color vision research: tutorial. *J. Opt. Soc. Am. A, JOSAA* **36**, 606–627 (2019).
4. CIE JTC 2. CIE 018:2019 The Basis of Physical Photometry, 3rd Edition. doi:10.25039/TR.018.2019.
5. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav Res* **51**, 195–203 (2019).
6. Needell, C. D. & Bainbridge, W. A. Embracing New Techniques in Deep Learning for Estimating Image Memorability. *Comput Brain Behav* **5**, 168–184 (2022).
7. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems* vol. 25 (Curran Associates, Inc., 2012).
8. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. Preprint at https://doi.org/10.48550/arXiv.1512.03385 (2015).
9. Khosla, A., Raju, A. S., Torralba, A. & Oliva, A. Understanding and Predicting Image Memorability at a Large Scale. in *2015 IEEE International Conference on Computer Vision (ICCV)* 2390–2398 (2015). doi:10.1109/ICCV.2015.275.
10. Goetschalckx, L. & Wagemans, J. MemCat: a new category-based image set quantified on memorability. *PeerJ* **7**, e8169 (2019).
11. Davis, T. M. & Bainbridge, W. A. Memory for artwork is predictable. *Proc Natl Acad Sci U S A* **120**, e2302389120 (2023).