Visual looming is a primitive for human emotion

Monica K. Thieu[1*], Vladislav Ayzenberg[1,2], Stella F. Lourenco[1], Philip A. Kragel[1*]

[1]Emory University, [2]University of Pennsylvania

* Please address correspondence to:

Monica Thieu
mthieu@emory.edu
Department of Psychology
Emory University
36 Eagle Row
Atlanta, GA 30322

Philip A. Kragel
pkragel@emory.edu
Department of Psychology
Emory University
36 Eagle Row
Atlanta, GA 30322
+1 (404) 727-3409

Abstract

Looming objects afford threat of collision across the animal kingdom. Defensive responses to looming and neural computations for looming detection are strikingly conserved across species. In mammals, information about rapidly approaching threats is conveyed from the retina to the midbrain superior colliculus, where variables that indicate the position and velocity of approach are computed to enable defensive behavior [1]. Neuroscientific theories posit that emotional feelings are based on representations in the midbrain [2], which are further elaborated in cortical systems [3]. However, how these computations relate to phenomenal experience in humans is largely unknown. Here, we show that a shallow convolutional neural network based on the *Drosophila* visual system predicts human behavioral and brain responses across development. In laboratory experiments using controlled visual stimuli, we find that this model explains defensive behaviors to looming objects in infants and superior colliculus responses to optical expansion in adults. Further, examining responses to a broader array of naturalistic video clips, we observe that representations of looming from this model predict self-reported emotion largely on the basis of subjective arousal. Our results illustrate how human emotions may be supported by species-general systems for survival in unpredictable environments.

Main Text

Emotions guide people to make sense of and react adaptively to the world around them. A hallmark of human emotion is the complexity of emotionally evocative situations and the varied ways in which individuals appraise them. Nevertheless, certain events consistently drive similar experiences across individuals. A spectator at a baseball game is likely to flinch in the face of an oncoming foul ball. A pedestrian might report feeling frightened after a speeding car cuts too close to them in the crosswalk. Even if emotional experience is ultimately highly personalized by a variety of developmental and cultural factors [4], some components of this experience are likely shared across people and across phylogeny. These building blocks of emotion are considered "primitives" in the sense that they are psychologically irreducible [5] and they have properties that are present across species [6]. Thus, to understand the nature and origins of human emotion, we must identify which features are shared across species and the mechanisms by which specific sensory inputs drive specific emotional states.

Humans are tuned to detect and react to certain classes of evolutionarily prepared stimuli, and threats to survival in particular [7,8]. Predators make up one such class of threats. For example, human observers—including infants and children—detect pictures of snakes faster than other objects [9–13]. Macaques also rapidly detect and learn to avoid snakes [14,15], two behaviors thought to be implemented in subcortical pathways through the superior colliculus and pulvinar [16]. Emotional expressions make up another such class of sensory signals indicative of threat. For example, fearful facial expressions are detected more rapidly than other expressions [17]. This heightened sensitivity may be subserved by the detection of specific visual features, like widened eyes, in the amygdala via similar inputs from the pulvinar nucleus [18]. However, not all animals are as sensitive to predatory snakes, or the wide-eyed facial expressions of conspecifics, suggesting these behaviors are unlikely to be supported by neural mechanisms that are shared across species.

One type of stimulus that is perceived as threatening and evokes defensive behavior across species is visual looming. As an object approaches the viewer, or *looms,* it tends to block light, and its edges expand optically. Additionally, if the object is on a collision course, its edges will expand *radially* in the observer's frame of reference. Rapidly approaching objects in the environment are almost invariably dangerous, like predators, or projectiles that may cause physical damage upon contact, and very few other types of environmental motion will create such a combination of visual features. Dark-shape radial expansion thus *affords* threat of collision to any animal that can detect it [19].

Many species of animals show defensive responses to looming stimuli that are subserved by similar neural pathways. Rapidly looming shadows elicit escape behaviors in animals including but not limited to insects, birds, rodents, and nonhuman primates [20–23]. Humans, as well, show defensive responses—when faced with physically looming objects, human infants and adults blink and flinch respectively [24–27]. Across mammals, detecting and responding to looming motion involves the superior colliculus, a midbrain structure whose neural organization and role in sensorimotor orienting is highly conserved across species [28]. The superior colliculus encodes looming via neurons in its deeper layers. Information about looming is used to coordinate defensive behavior via projections to subcortical structures including the periaqueductal grey, ventral tegmental area, and the thalamus [29]. The computations involved in detecting and responding to visually looming threats are comparable across vertebrates [1], suggesting they may produce a "central emotion state" that is a building block of emotion [6]. This

conservation suggests models of looming detection from nonhuman animal studies can be applied to predict human responses to similar stimuli.

Although the brain basis of looming detection is increasingly understood across species, it remains unclear how visual looming contributes to affective experience in humans. Looming is such a strong threat cue that one can readily imagine one's emotional response to, say, seeing a ball hurtle toward one's head, even if the ball does not actually make contact. Even though looming is well-established as an aversive and arousing experience [30,31], we still lack a mechanistic understanding of how looming is detected by human perceptual systems, and how perceptual representations of looming relate to subjective experience. To this end, we use a shallow convolutional neural network based on the architecture of the *Drosophila* visual system [32] as a computationally explicit model of collision detection.

We hypothesize that visual looming contributes to human emotional experience via computations that are conserved across species and only require information available in the optical array [19]. If this is the case, then a species-general convolutional neural network for collision detection should predict brain activity, defensive behavior, and subjective experience in humans, because it contains representations related to "primitive" dimensions of arousal and threat. To test these predictions, we analyze human adult fMRI responses to dynamic videos [33] to assess whether BOLD signal in the superior colliculus encodes representations of looming from the artificial neural network. Next, we analyze looming stimuli that induce defensive blinking in human infants to assess whether the same representations of looming predict defensive behavior. Finally, we analyze naturalistic videos that evoke twenty distinct varieties of emotional experience to assess whether representations of looming predict self-reported emotion, and whether those representations relate to valence and arousal, which together are thought to comprise a building block of emotional experience [34,35]. Through these analyses, we test which aspects of human affective experience can be produced by computations implemented in the perceptual systems of multiple species.

We adapted a model of *Drosophila* collision detection to relate the looming component of naturalistic videos to human affective responses. We implemented a pre-trained shallow convolutional neural network with connections constrained by the connectivity of loom-sensitive *Drosophila* LPLC2 neurons [32], directly inputting the pre-trained filter for a single LPLC2 "neuron" as the kernel (Figure 1A). The network takes sequences of optical flow as input, with four channels per frame–one for each of the cardinal directions of optical flow. These channels are convolved with a characteristic radial outward motion filter across each frame, producing a two-dimensional spatial representation of looming. The looming representation is summed across units to produce a framewise estimate of collision probability over the sequence of visual inputs. We used the activations in the single convolutional layer (i.e., representations of looming) to predict human brain responses in superior colliculus, infant defensive blinking, and self-reported emotion evoked by naturalistic videos.

Visual looming is encoded in the human superior colliculus

First, we tested whether variables used to predict imminent collision in the shallow convolutional neural network are encoded in human superior colliculus activity. We fit encoding models [36] of looming motion to predict fMRI signal acquired as participants (N = 15) viewed dynamic visual stimuli used for retinotopic mapping [33] (see Methods). The visual stimuli included four types of motion: clockwise and counterclockwise sweeping wedges in addition to contracting and expanding rings. Because expanding rings involve symmetric radial expansion, which is a hallmark of looming and activates specific neural

populations in deep layers of the superior colliculus [37], responses to this stimulus should be best explained by encoding models utilizing features that are useful for detecting imminent collision.

Accordingly, to test for looming specificity, we compared the performance between two encoding models: a stimulus-general model that identifies mappings between representations of looming and human brain activity using responses to all four stimulus types, and an expansion-specific model that identifies mappings using only responses to optical expansion. If the human superior colliculus responses encode visual looming, then the encoding model trained on optical expansion alone should outperform the stimulus-general model.

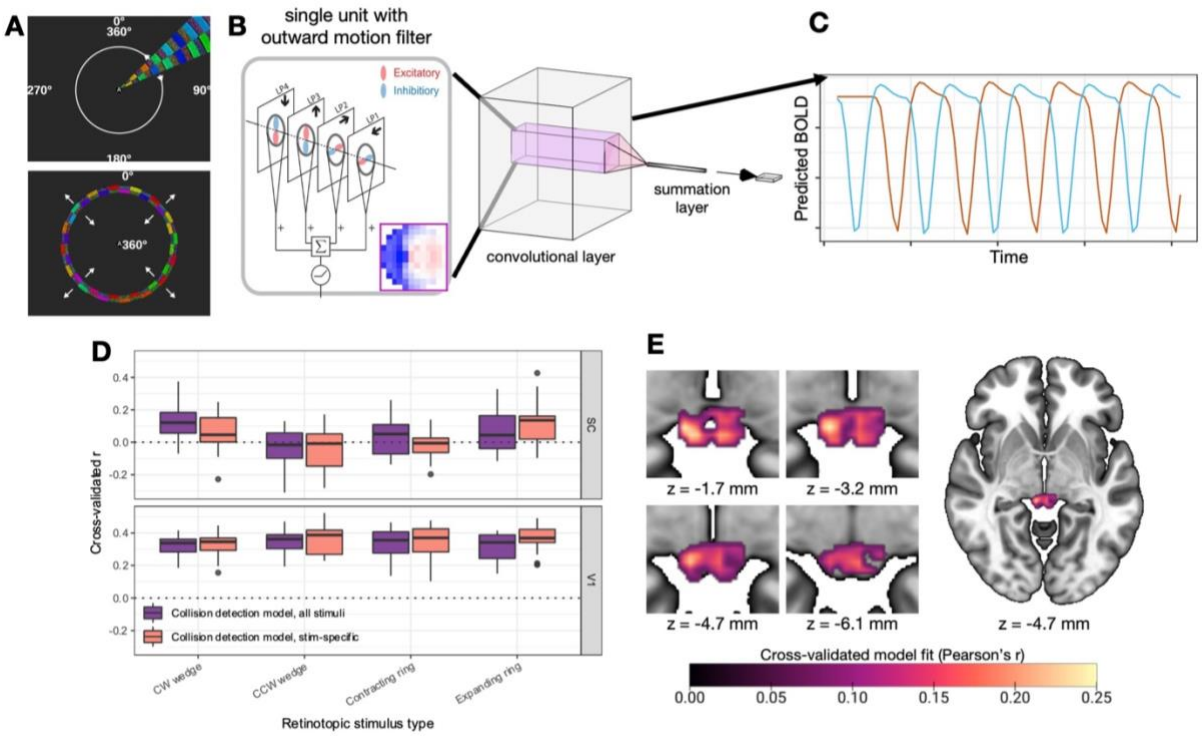

Figure 1. Visual looming is encoded in the human superior colliculus. (A) We computed optical flow from dynamic retinotopic mapping stimuli featuring clockwise and counterclockwise sweeping wedges (top) and contracting and expanding rings (bottom). **(B)** We passed the optical flow through a shallow convolutional neural network originally trained to detect imminent collision. The pre-trained convolutional units (left) filter each frame for outward motion in the four cardinal directions, and output a matrix of activations corresponding to the timecourse of looming motion at various points in the visual field. Figure adapted with permission from reference [32]. **(C)** We convolved these timecourses with the canonical hemodynamic response function to generate a multivariate encoding model to predict looming-related BOLD signal. Example timecourses are shown for a 5-cycle run of the expanding ring stimulus from two units at the center (blue) and periphery (orange) of the visual field. Units at the center tend to peak in activation early in the cycle, when the ring is in the center of the visual field, and units at the periphery tend to peak later in the cycle, when the ring has expanded. **(D)** We assessed model performance using leave-one-subject-out cross-validated Pearson's r between encoding model-predicted bold and observed BOLD. Boxplots show median and interquartile range of the model fit estimates from each cross-validation fold. The expansion-specific model of superior colliculus activity (pink, far right, upper panel) outperforms a stimulus-general model on

the same data. CW = clockwise. CCW = counterclockwise. **(E)** Voxelwise activity was well-explained by the expansion-specific model across our superior colliculus ROI.

We found that the expansion-specific encoding model predicted BOLD responses in the superior colliculus (leave-one-subject-out cross-validated $r = .119$, $p < .001$, permutation test; Figure 1D), and that it outperformed a stimulus-general encoding model that was not specific to expanding motion ($\Delta r = .046$, $p = .020$, permutation test). Critically, the expansion-specific model performed equally or worse than the general model when predicting responses to sweeping wedges and contracting rings ($\Delta r = .092$, $p < .001$, permutation test). On the other hand, we found that V1 BOLD responses, which should not be predominantly driven by any particular stimulus, were explained well by the stimulus-general model (leave-one-subject-out cross-validated $r = .332$, $p < .001$, permutation test; Figure 1D). The expansion-specific encoding model also slightly outperformed other stimulus-specific models when adjusting for stimulus-general encoding model fit, with a notably smaller effect size than models of superior colliculus activity ($\Delta r = .043$, $p = .002$; Figure 1D). These results confirm that patterns of BOLD activity in V1 contain information about both the general location and motion of visual stimulation, whereas patterns of BOLD activity in the human superior colliculus encode representations of looming motion that predict defensive behavior across species.

Representations of looming predict defensive blinking in infants

To investigate whether the shallow convolutional network can characterize the behavioral affordances of looming threats, we evaluated whether it predicts defensive blinking in human infants. Infants develop a propensity to blink in the face of looming stimuli beginning at 4-6 months [26]. Defensive blinking is selective to impending collision, and thus likely driven by looming computations like the ones modeled by our shallow neural network. If the shallow neural network contains representations of looming similar to those used by newborn infants, then infants' tendency to blink while viewing looming objects should be related to model-estimated collision probability on each frame.

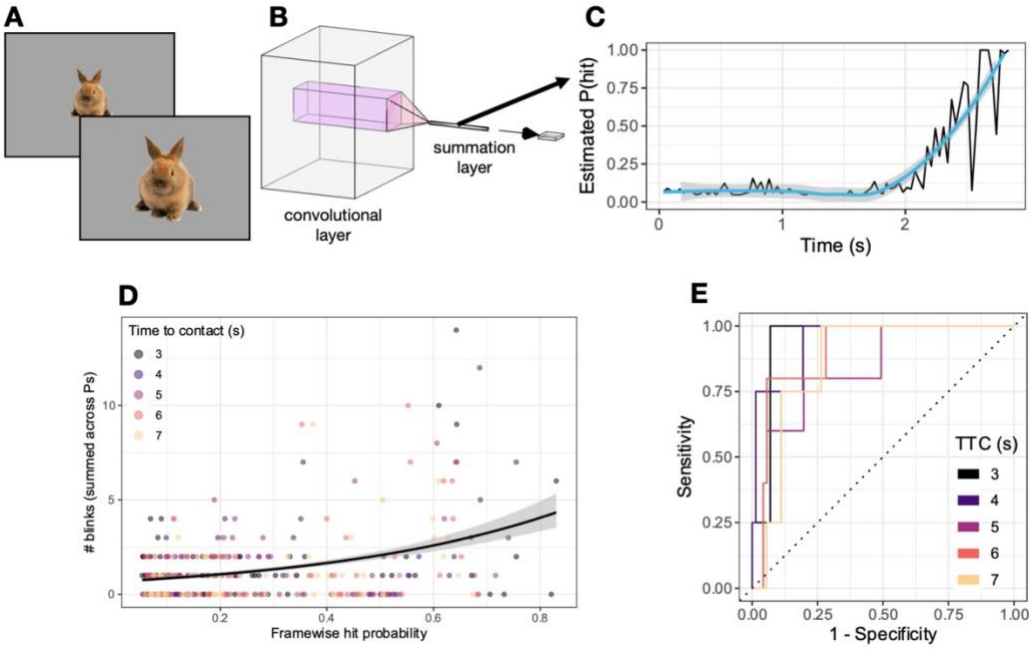

Figure 2. Representations of looming predict defensive blinking to looming objects in infants. (A) Infants viewed videos of looming objects generated by radially expanding static images over time to simulate the appearance of approach motion. **(B)** We passed the optical flow from these videos through the neural network for collision detection**,** which rectifies and sums unit activations and then applies a softmax activation function to estimate collision probability on each frame. **(C)** Extracted collision probabilities for one representative video. **(D** Videos with varying apparent times-to-contact showed that greater looming collision probability was associated with increased blinking on a given frame (Poisson regression). **(E**) Receiver operating characteristic curves showing separability of "high-blink" (>= 5 blinks) and "low-blink" (<5 blinks) frames.

Analyzing defensive blinking as infants viewed visually looming objects (Figure S2), we found that collision probability predicted blink count across all frames (*beta* = .221, *SE* = .020, Poisson regression, $p < .001$, permutation test; Figure 2D). To quantify the strength of this relationship, we leveraged the neural network's stronger activation to faster-approaching stimuli to test whether infants are similarly sensitive to the approach velocity of looming stimuli. We found that model-based representations of looming discriminated frames containing five or more total eye-blinks (which were generally at the end of the video clip, see Methods) from those that contained fewer than 5 eye-blinks (area under the ROC curve (AUROC) = .902, *SE* = .025, $p < .001$, permutation test), with greater discriminability with increasing object speeds (Kendall's $\tau = -.657$, $p = .046$, permutation test; Figure 2E). These observations confirm that species-general computations of optical expansion are sufficient to predict velocity-sensitive human defensive responses to dynamic looming stimuli.

Representations of looming predict subjective emotion elicited by naturalistic videos

Next, we investigated whether computations for detecting looming threats also predict normative ratings of emotional videos. Although many sources of information can inform subjective experience, here we tested whether visual input that affords imminent collision predicts variation in self-reported emotion. To this end, we trained a 20-way linear discriminant analysis classifier to identify the consensus emotion category of 1315 videos [38] from their two-dimensional looming representations. We tested this classifier on 332 held-out videos from the same database and found that representations of looming predicted the top consensus emotion category above chance (16.9%, SE = 2.1%, chance = 13.0%, $p = .010$, permutation test; Figure 3D). The AUROC was .538 (chance = .5, SE = .024, $p = .024$, permutation test), showing that looming information could discriminate between a subset of emotion classes.

Information about object motion and stable visual properties (e.g., texture, shape, color) are processed by distinct neural pathways [39], suggesting that information from these two sources may uniquely contribute to emotional experience. To test whether looming information predicts video emotion categories independently from information related to object quality, we compared the performance of the shallow neural network to a deep convolutional network that categorizes emotional situations based on the static content of individual video frames [40]. The ability of the looming model and the static visual model to classify emotion categories were uncorrelated (Kendall's $\tau = -.189$, $p = .122$, permutation test). For example, the looming model was relatively better at discriminating videos labeled as "fear" (looming model: AUROC = .638, SE = .066, 2nd out of 20; static model: AUROC = .727, SE = .064, 13th out of 20; Figure S2) and "surprise" (looming model: AUROC = .593, SE = .106, 5th out of 20; static model: AUROC = .565, SE = .137, 18th out of 20; Figure S2). Meanwhile, the static model was relatively better at discriminating videos labeled as "craving" (looming model: AUROC = .345, SE = .091, 19th out of 20; static model: AUROC = .969, SE = .017, 1st out of 20; Figure S2) and "sexual desire" (looming model: AUROC =

.465, SE = .074, 13th out of 20; static model: AUROC = .943, SE = .018, 2nd out of 20; Figure S2). These differences suggest that some emotion categories are characterized by the presence of looming motion, whereas other categories are better characterized by the presence of specific visual features, irrespective of how those features move in the environment.

Finally, we investigated how the representation of looming in the shallow convolutional neural network relates to different emotion categories. To distinguish prevalent psychological accounts of looming [19,34,41], we compared the similarity of predictions made when attempting to classify patterns of optical flow into 20 emotion categories (Figure S3) to the similarity of self-report ratings of fear, valence, and arousal (representational similarity analysis [42]; see Figures S4-5). We also included the similarity of predictions from the deep convolutional neural network as a covariate (Figure S6), to assess relationships between model-based representations of looming and self-reported experience, independent of static visual features.

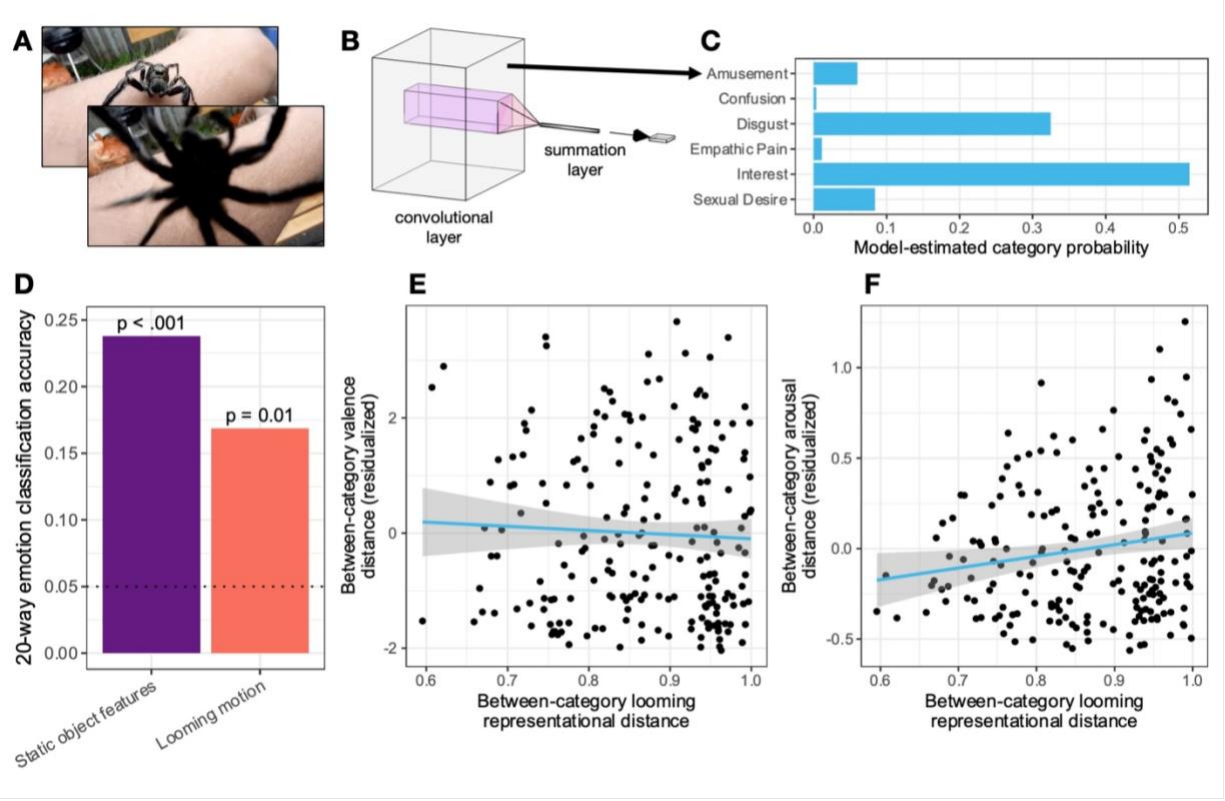

Figure 3. Representations of looming predict subjective emotion evoked by naturalistic videos in adults. (A) Participants viewed short video clips depicting a variety of situations. Frames are shown from a stimulus with apparent looming motion. **(B)** We passed the optical flow from these videos through the same convolutional neural network and extracted unit activations from the convolutional layer. **(C)** We trained a 20-way linear discriminant classifier to predict the normative emotion category of each video from its looming activations. **(D)** The collision detection model classified videos into 20 consensus emotion categories above chance (right bar). A deep convolutional neural network that uses static visual features associated with different emotion categories also performed well (left bar). **(E)** Distance between emotion categories in the collision detection emotion classifier is unrelated to subjective valence, after adjusting for information from the static object quality-based emotion classifier. **(F)** Distance between emotion categories in the collision detection classifier

is associated with subjective arousal, after adjusting for information from the static object quality-based classifier.

This analysis revealed that the similarity of emotion categories in the looming-based classifier did not correlate with that of subjective fear (partial $r = .110$, $p = .121$, permutation test). The representational similarity of emotions based on looming positively correlated with arousal (partial $r = .169$, $p = .015$, permutation test; Figure 3F) but not valence (partial $r = .047$, $p = .495$, permutation test; Figure 3F). Meanwhile, representational similarity based on static visual features positively correlated with that of subjective fear (partial $r = .184$, $p = .008$, permutation test), arousal (partial $r = .237$, $p < .001$, permutation test) and valence (partial $r = .174$, $p = .012$, permutation test), conceptually replicating prior assessments of still frames from the same dataset [40]. These findings suggest that in this set of brief naturalistic videos, representations of looming motion that facilitate the detection of imminent collision are primarily related to subjective arousal, whereas representations of static visual features are more directly linked to valence and the experience of specific emotion categories.

Discussion

Our results support the view that emotion is functional; it flexibly guides cognition and enables organisms to respond to urgent needs in unpredictable environments [43]. We demonstrate how an incredibly simple network architecture can have broad explanatory power, accounting for different neurobehavioral measures across the lifespan. Recent advances using goal-driven optimization with much more complex architectures (on the order of 10^7 more parameters) to characterize cortical systems involved in object recognition, speech perception, and language processing [44–46] has been based on the idea that large, overparameterized models are necessary to explain the human mind. The present findings stand in contrast to this approach, illustrating how a much simpler architecture trained with the right objective function—a computational primitive—characterize multiple aspects of human behavior that are not explained by more complex models of cortical brain systems [40].

Studies focused on human language and self-reported experience identify valence (pleasantness or unpleasantness) as the dimension that best predicts the semantic structure of emotion across cultures [47]. Experience-sampling suggests that adults organize their emotions primarily using valence [35], and developmental studies further show that infants and children first distinguish facial expressions and linguistic concepts using valence [48,49]. We found that computations supporting a species-general behavior uncover information predominantly related to subjective arousal, suggesting that primitive aspects of phenomenal experience may be implemented at the level of the human midbrain [2]. More generally, our findings caution against the assumption that certain stimuli which evoke defensive behaviors produce experiences that resemble prototypical instances of fear in adults, because the computations that underlying these behaviors do not strongly predict subjective valence or fear in a broader array of naturalistic stimuli.

Our work validates nonhuman animal research examining the neural mechanisms of looming detection. Because the artificial neural network we studied is sensitive to velocity and predicts imminent collision similarly to the computation of η in Drosophila LPLC2 neurons [32,50], our fMRI findings suggest that this optical parameter is encoded in the human superior colliculus as well. Because neuronal populations encoding η fire maximally just prior to collision, they are particularly strong candidates for mediating

subsequent escape behavior. Escape vigor is calculated by thresholding this population activity in downstream regions like the insect lobula giant movement detector [51] and the mammalian dorsal periaqueductal grey [52]. Analogously, future work in humans can investigate the flow of continuous looming threat from superior colliculus to interconnected regions that coordinate defensive behavior.

We showcase the power of characterizing how human emotions may arise from ancient computations conserved across species. Our observations provide a sketch of what a complete understanding of emotion might look like from a neurocomputational perspective. Precisely characterizing species-general central emotion states [6] by modeling how environmental and social affordances shape behavior (e.g., active escape *coupled with* the looming stimuli that elicit it) will likely explain a substantial portion of human emotion. Developing models that can faithfully perform other tasks involving language, music, and complex social behaviors that are uniquely human will also likely be necessary. By shifting the focus from a small number of apparently simple, interpretable dimensions to computationally explicit models that match the complexity of the brain [53], this approach promises to yield fundamentally new insights into the origins and nature of emotion.

Materials & methods

Implementation of the shallow convolutional neural network

We implemented a shallow neural network model originally built to model the Drosophila LPLC2 pathway and trained to identify whether dynamic stimuli are on a collision course with the viewer [32]. The network takes in a 4D timecourse of visual motion in each of the 4 cardinal directions. The network has two layers that operate on each frame of the timeseries: one convolutional layer, which, once trained, passes a 12 × 12 px outward motion filter over the visual field to generate a 256-unit representation of looming, and one summation layer, which rectifies, sums, and applies a softmax activation function to estimate looming collision probability for that frame.

For each of the studies reported below, we first resized the study's stimuli to 132 × 132 px to yield 256 convolutional units given the filter size and stride parameters. We then estimated each stimulus' optical flow using the Farneback algorithm as implemented by OpenCV [54,55] and re-cast the optical flow from 2D (positive/negative motion in the x and y directions) to 4D (positive motion in each of the cardinal directions, hereafter referred to as cardinal flow) in accordance with the model.

We then adapted the pre-trained collision detection model from operating on fly-like to human-like vision, instantiating it as a 2D convolutional neural network in PyTorch [56] that passes the pre-trained 12 × 12 px outward motion filter over the optical flow from a human-watchable video stimulus, with 11 px stride and 0 px padding, to replicate the unit-to-unit visual field overlap from the original fly-like model. We left the summation layer identical to the original model. Finally, we passed each stimulus' cardinal flow through the modified collision detection model and extracted representations of looming at various stages of the model to map onto human responses (described further for each study below).

Study 1: Retinotopic fMRI study

Overview. In Study 1, we tested whether looming representations in our model were encoded in human superior colliculus BOLD activity. We leveraged whole-brain fMRI responses to dynamic visual stimuli used for retinotopic mapping to maximize potential looming-related variance in superior colliculus activity. We hypothesized that BOLD responses to visual stimuli would be driven by two types of neural populations: retinotopically organized populations in superficial layers that respond irrespective of motion direction [57] and populations in intermediate and deep layers of the colliculus that respond primarily to expanding radial motion [58]. We tested this hypothesis by fitting multivariate encoding models to predict patterns of colliculus response using the shallow convolutional neural network for collision detection as a feature extractor. If the human superior colliculus contains neural populations that code for visual looming, and they are engaged by the retinotopic videos, then encoding model performance should be the highest on models trained and tested specifically on video stimuli that include optical expansion.

Experimental paradigm and stimuli. We used a previously published dataset of retinotopic mapping fMRI scans collected on 15 healthy adult participants [33]. Participants were scanned while viewing four types of dynamic retinotopy stimuli: clockwise and counterclockwise sweeping wedges, and contracting and expanding rings. The stimuli cycled across the visual field with a period of 32 s, with five repetitions per run, with each run lasting 3 minutes.

MRI preprocessing. fMRI data were preprocessed using SPM12 in MATLAB [59,60]. Images were first realigned to the first image of the series using a six parameter, rigid-body transformation [61]. The

realigned images were then normalized to MNI152 space using a 12-parameter affine transformation followed by nonlinear deformations using a three-dimensional discrete cosine transform basis set, as implemented in SPM [62,63]. No additional smoothing was applied to the normalized images. Normalized images were subsequently temporally bandpass filtered with cutoff frequencies centered around the stimulus frequency (.667/32 and 2/32 Hz).

Measurements. We extracted preprocessed BOLD timeseries from a hand-drawn ROI of the superior colliculus [64,65], as well as from an ROI of V1 [66] as a positive control.

Modeling. We passed sequences of cardinal flow from each retinotopic mapping stimulus through the convolutional layer of the collision detection model. We then convolved the timecourse of units in the shallow convolutional neural network to each of the retinotopy stimuli with the SPM double-gamma hemodynamic response function to generate a multivariate encoding model of looming-related BOLD signal. We applied partial least-squares (PLS) regression, implemented through the mixOmics and tidymodels packages in R [67–69], to map our looming-predicted BOLD onto observed multivariate BOLD from each ROI separately. We trained the PLS multivariate encoding model on data from 14 participants and then assessed model fit as the Pearson correlation between PLS-predicted BOLD and observed BOLD in the last held-out participant. We cross-validated model fit in a leave-one-subject-out manner by repeating this process for every participant and averaging across repetitions [70].

Because the collision detection model contains units that tile the visual field, the resulting BOLD encoding model encodes both retinotopic responses and responses to looming motion. Accordingly, to test for looming specificity, we compared performance between two types of encoding models: a stimulus-general model, with the PLS mapping trained on data from all four stimulus types, and stimulus-specific models, with the PLS mapping trained separately on data from each stimulus type. We expected the stimulus-specific model trained on expanding ring motion would predict superior colliculus responses more so than other stimulus-specific models, or the stimulus-general model.

We generated block permutation distributions against which to compare the model fit correlations by randomizing TRs of observed BOLD within each stimulus cycle to preserve the autocorrelation structure of the data [71]. We then re-estimated each shuffled model fit correlation over 5,000 iterations to generate p-values for inference.

Study 2: Infant behavioral study

Overview. In Study 2, we tested whether looming representations in our model could predict infant defensive blinking in response to looming stimuli.

Participants. A total of 62 healthy infants participated in this study. Of the 62 infants, four infants looked less than 35% of the (total) trial durations and, thus, were excluded from subsequent analyses. An additional 12 infants failed to complete the study due to fussiness or technical difficulties, leaving 58 infants in the final sample (*range* = 6.2-11.7 months, *M* = 8.7 months; 22 boys and 36 girls). Parents provided written informed consent on behalf of their infants. All procedures were approved by the Institutional Review Board at Emory University.

Procedure and stimuli. Infants were tested individually in a dimly lit, soundproof room. Each infant sat in a highchair or on his/her parent's lap at a distance of approximately 60 cm from a large projection screen (92.5 × 67.5 cm). Parents were instructed to keep their eyes closed and to refrain from

interacting with their infants during the study, except for soothing them if they became fussy. Stimuli were videos of a looming two-dimensional image, which were rear-projected onto the screen at eye-level to the infant. Each infant's face was recorded for later coding using a concealed camcorder placed just under the projection screen. Video feed was transmitted directly to a computer in an adjoining room where an experimenter monitored the session remotely.

Images in each of the videos were of individual animals (snakes, spiders, butterflies, and rabbits; two of each type); see figure below. Images were selected from an Internet search for their high quality and to match roughly in color and brightness. Images were cropped, resized, and presented against a uniform gray background using Adobe Photoshop CS5 [72]. Looming videos were created in MATLAB by manipulating the rate of expansion of the image size.

Each trial was experimenter controlled, beginning with a centrally presented attention-getter (e.g., swirling star; randomly selected across trials) that played until infants oriented to the screen. A looming video immediately followed. Each video began with two-dimensional image that expanded to a maximum size of 75° × 59° (visual angle). There was a 1 s inter-trial interval (ITI) consisting of a gray screen. Videos were created such that the virtual animal approached the infant at one of six velocities, indicating times-to-contact of 3, 4, 5, 6, 7, or 8 s. Velocity was negatively correlated with approach time, such that as approach time increased, the velocity of the virtual object decreased. Infants were presented with a total of 48 trials (randomized).

Video coding. High quality videos of each infant were saved digitally. Video frames were coded at 33.33 ms intervals by observers blind to the stimuli presented to infants. All videos were coded by one observer for blinks (and total looking time) on each trial. Eye closures were counted as blinks if the lids of the opened eyes covered at least half of the exposed eye surface [73]. Incomplete eye closures associated with large head turns were not counted as blinks. Also not counted as blinks were eye closures associated with yawns, sneezes, coughs, and hand movements to or near the face or mouth. A second observer coded a random sample of videos (20%) to assess reliability. Inter-observer reliability was high for the coding of both blinks and looking times (rs > 0.9).

Measurements. For each looming video stimulus presented to the infants, we summed the total number of blinks made by all infants on each coded frame to generate one timecourse of blink counts per video stimulus. We then further summed the blink count timecourses for each video of a given time-to-contact duration to generate one timecourse of total blink counts per time-to-contact condition (Figure S1).

Modeling. We extracted the cardinal optical flow for each looming video stimulus at a frame rate of 33.33 ms/frame, and then passed the flow videos through the convolutional and summation layers of the collision detection model to generate a 1D timecourse of estimated collision probability for each stimulus. We then averaged the timecourses for each video of a given time-to-contact duration to generate one timecourse of looming collision probability per time-to-contact duration.

Then, we used Poisson regression to predict framewise blink counts as a function of framewise collision probability and condition-wise time-to-contact. We generated a permutation distribution against which to compare the coefficient for collision probability by randomizing blink counts across all trials. We then re-fit the Poisson regression and extracted the shuffled coefficient over 10,000 iterations to generate p-values for inference.

Finally, we examined the potentially threshold-like relationship between blink counts and collision probability by using collision probability to classify frames as "high-blink" (5 or more blinks across infants/stimuli on that frame, to isolate trials where blinks were most likely to be defensive) or "low-blink" (fewer than 5 blinks). We calculated the area under the receiver operating curve (AUROC) both overall and as a function of time-to-contact condition, using tools implemented in the tidymodels family of R packages [68]. We evaluated whether AUROC varied with time to collision by calculating Kendall's τ between the observed rank-ordering of times-to-contact based on AUROC (highest to lowest) and duration (3 s to 7 s). We generated a non-parametric sampling distribution for overall AUROC by bootstrap resampling and re-calculating AUROC over 10,000 iterations. We also generated a permuted distribution against which to compare the observed AUROC by randomizing binarized blink counts across all trials and re-estimating AUROC over 10,000 iterations. Similarly, we generated a block permutation distribution against which to compare the observed Kendall rank correlation between time-to-contact and AUROC by randomizing binarized blink count within each time-to-contact condition. We then re-estimated the shuffled AUROC for each time-to-contact and re-calculated Kendall's τ over 10,000 iterations to generate *p*-values for inference.

Study 3: Adult behavioral study

Overview. In Study 3, we tested whether looming representations in our model could predict normative self-report affect ratings in response to short, naturalistic videos.

Stimuli and behavioral measurements. We used a previously published subset of short, naturalistic videos and normative emotion ratings [38]. Each video was rated by approximately 10 raters (range = [9, 17]), each of whom reported the categorical emotions elicited by the video, as well as 9-point valence and arousal ratings. For each video, we took its most frequently selected categorical emotion label, and its mean valence and arousal ratings. Videos spanned 20 consensus emotion categories.

Modeling. We resampled each video stimulus to a standard frame rate of 10 fps and passed the cardinal flow from each video stimulus through the convolutional layer of the collision detection model to yield 256 timecourses of activations per video. Next, we flattened each video's looming representation along the time dimension. The original looming model tends to *increase* activation over time for "hit" stimuli as the stimuli approach the viewer and activate an increasing number of units across the visual field. Accordingly, we assumed that stronger looming activations would have a more positive slope over time. We calculated the linear slope of each unit's timecourse over time, generating a looming representation of 256 unit activation slopes per video. We then applied linear discriminant analysis, implemented through the MASS and the tidymodels packages in R, to classify each video's consensus emotion category (out of 20) using its 256 looming activation slopes. We trained the linear discriminant classifier using a prior training split of 1315 videos [40]. All model performance statistics are reported as evaluated on the associated prior held-out testing split of 332 videos.

We compared the emotion classification performance of the looming model to the performance of a deep convolutional neural network originally trained to classify stimulus-elicited emotions based on their static image features [40]. Because that model was originally used to identify the emotion categories of individual video frames, we calculated video-wise category predictions by averaging each of the 20 emotion class probabilities across each frame of the video and taking the emotion category with the highest across-video average probability. We generated non-parametric sampling distributions for our statistics by bootstrapping and re-calculating classification accuracy, over 10,000 iterations. We also

generated non-parametric null distributions against which to compare classification accuracies by permuting the consensus emotion category labels across videos and re-calculating shuffled classification accuracy, over 10,000 iterations. Finally, we generated a permutation distribution against which to compare Kendall's τ for category rankings by model AUROC by randomizing consensus emotion category label the across videos. We then re-estimated shuffled category-specific AUROCs for both the looming model and the static image model, and re-calculated a shuffled Kendall's τ over 10,000 iterations to generate p-values for inference.

We used representational similarity analysis [42] to assess whether the representations learned by the emotion classification models encoded information consistent with valence and/or arousal. For both the looming motion-based and static visual feature-based classifiers, we calculated the representational distance between every pair of emotion categories. For a given emotion classification model and pair of emotion categories, we calculated the distance as 1 minus the average pairwise Pearson correlation between the 20 class probabilities for any two videos from those two emotion categories. We then used linear regression to predict between-category distances in mean valence ratings from distances from both convolutional networks, allowing us to assess the independent contributions of information gleaned from optical flow and static visual features. From this regression, we estimated the partial correlation coefficients that identify the relationship between representations of looming and valence (accounting for static visual features), and between representations of static visual features and valence (accounting for looming). We conducted similar regressions using mean ratings of arousal and fear, and extracted partial correlation coefficients using the same approach. We generated permutation distributions against which to compare these partial correlation coefficients [74,75], calculating randomized partial correlation coefficients over 10,000 iterations to generate p-values for inference.

481 ## References

482 1. Peek, M. Y. & Card, G. M. Comparative approaches to escape. *Curr. Opin. Neurobiol.* **41**, 167–173
483 (2016).
484 2. Damasio, A. & Carvalho, G. B. The nature of feelings: evolutionary and neurobiological origins. *Nat.*
485 *Rev. Neurosci.* **14**, 143–152 (2013).
486 3. LeDoux, J. E. & Brown, R. A higher-order theory of emotional consciousness. *Proc. Natl. Acad. Sci.*
487 **114**, E2016–E2025 (2017).
488 4. Lindquist, K. A., Jackson, J. C., Leshin, J., Satpute, A. B. & Gendron, M. The cultural evolution of
489 emotion. *Nat. Rev. Psychol.* **1**, 669–681 (2022).
490 5. Barrett, L. F. & Bliss-Moreau, E. Affect as a Psychological Primitive. *Adv. Exp. Soc. Psychol.* **41**, 167–
491 218 (2009).
492 6. Anderson, D. J. & Adolphs, R. A Framework for Studying Emotions across Species. *Cell* **157**, 187–200
493 (2014).
494 7. Cosmides, L. & Tooby, J. Evolutionary Psychology and the Emotions. in *Handbook of emotions* (eds.
495 Lewis, M. & Haviland-Jones, J. M.) (Guilford, 2000).
496 8. LoBue, V. & Rakison, D. H. What we fear most: A developmental advantage for threat-relevant
497 stimuli. *Dev. Rev.* **33**, 285–303 (2013).
498 9. Alvarez, L. C. & Pipitone, R. N. Replication of LoBue & DeLoache (2008, PS, Study 3). (2013).
499 10. Lazarević, L. B. *et al.* Many Labs 5: Registered Replication of LoBue and DeLoache (2008). *Adv.*
500 *Methods Pract. Psychol. Sci.* **3**, 377–386 (2020).
501 11. Bertels, J. *et al.* Snakes elicit specific neural responses in the human infant brain. *Sci. Rep.* **10**, 7443
502 (2020).
503 12. LoBue, V. & DeLoache, J. S. Detecting the Snake in the Grass: Attention to Fear-Relevant Stimuli by
504 Adults and Young Children. *Psychol. Sci.* **19**, 284–289 (2008).

13. Öhman, A., Flykt, A. & Esteves, F. Emotion drives attention: Detecting the snake in the grass. *J. Exp. Psychol. Gen.* **130**, 466–478 (2001).
14. Shibasaki, M. & Kawai, N. Rapid Detection of Snakes by Japanese Monkeys (Macaca fuscata): An Evolutionarily Predisposed Visual System. *J. Comp. Psychol. Wash. DC 1983* **123**, 131–5 (2009).
15. Öhman, A. & Mineka, S. The Malicious Serpent: Snakes as a Prototypical Stimulus for an Evolved Module of Fear. *Curr. Dir. Psychol. Sci.* **12**, 5–9 (2003).
16. Le, Q. V. *et al.* Pulvinar neurons reveal neurobiological evidence of past selection for rapid detection of snakes. *Proc. Natl. Acad. Sci.* **110**, 19000–19005 (2013).
17. Adolphs, R. Fear, faces, and the human amygdala. *Curr. Opin. Neurobiol.* **18**, 166–172 (2008).
18. Barrett, L. F. Seeing Fear: It's All in the Eyes? *Trends Neurosci.* **41**, 559–563 (2018).
19. Gibson, J. J. The Theory of Affordances. in *The Ecological Approach to Visual Perception: Classic Edition* 119–135 (Psychology Press, 2014). doi:10.4324/9781315740218.
20. Card, G. M. Escape behaviors in insects. *Curr. Opin. Neurobiol.* **22**, 180–186 (2012).
21. Schiff, W., Caviness, J. A. & Gibson, J. J. Persistent Fear Responses in Rhesus Monkeys to the Optical Stimulus of 'Looming'. *Science* **136**, 982–983 (1962).
22. Wang, Y. & Frost, B. J. Time to collision is signalled by neurons in the nucleus rotundus of pigeons. *Nature* **356**, 236–238 (1992).
23. Yilmaz, M. & Meister, M. Rapid Innate Defensive Responses of Mice to Looming Visual Stimuli. *Curr. Biol.* **23**, 2011–2015 (2013).
24. Ball, W. & Tronick, E. Infant Responses to Impending Collision: Optical and Real. *Science* **171**, 818–820 (1971).
25. King, S., Dykeman, C., Redgrave, P. & Dean, P. Use of a Distracting Task to Obtain Defensive Head Movements to Looming Visual Stimuli by Human Adults in a Laboratory Setting. *Perception* **21**, 245–59 (1992).

26. Yonas, A. *et al.* Development of sensitivity to information for impending collision. *Percept. Psychophys.* **21**, 97–104 (1977).
27. Kayed, N. S. & van der Meer, A. Timing strategies used in defensive blinking to optical collisions in 5- to 7-month-old infants. *Infant Behav. Dev.* **23**, 253–270 (2000).
28. Basso, M. A. & May, P. J. Circuits for Action and Cognition: A View from the Superior Colliculus. *Annu. Rev. Vis. Sci.* **3**, 197–226 (2017).
29. Liu, X. *et al.* The Superior Colliculus: Cell Types, Connectivity, and Behavior. *Neurosci. Bull.* **38**, 1519–1540 (2022).
30. Bach, D. R., Neuhoff, J. G., Perrig, W. & Seifritz, E. Looming sounds as warning signals: The function of motion cues. *Int. J. Psychophysiol.* **74**, 28–33 (2009).
31. Riskind, J. H. & Maddux, J. E. Loomingness, Helplessness, and Fearfulness: An Integration of Harm-Looming and Self-Efficacy Models of Fear. *J. Soc. Clin. Psychol.* **12**, 73–89 (1993).
32. Zhou, B., Li, Z., Kim, S., Lafferty, J. & Clark, D. A. Shallow neural networks trained to detect collisions recover features of visual loom-selective neurons. *eLife* **11**, e72067 (2022).
33. Sengupta, A. *et al.* A studyforrest extension, retinotopic mapping and localization of higher visual areas. *Sci. Data* **3**, 160093 (2016).
34. Russell, J. A. Core affect and the psychological construction of emotion. *Psychol. Rev.* **110**, 145–172 (2003).
35. Barrett, L. F. Valence is a basic building block of emotional life. *J. Res. Personal.* **40**, 35–55 (2006).
36. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).
37. Lee, K. H., Tran, A., Turan, Z. & Meister, M. The sifting of visual information in the superior colliculus. *eLife* **9**, e50678 (2020).

38. Cowen, A. S. & Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7900–E7909 (2017).

39. Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G. & Mishkin, M. The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* **17**, 26–49 (2013).

40. Kragel, P. A., Reddan, M. C., LaBar, K. S. & Wager, T. D. Emotion schemas are embedded in the human visual system. *Sci. Adv.* **5**, eaaw4358 (2019).

41. Scherer, K. R. Profiles of Emotion-antecedent Appraisal: Testing Theoretical Predictions across Cultures. *Cogn. Emot.* **11**, 113–150 (1997).

42. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, (2008).

43. Simon, H. A. Motivational and emotional controls of cognition. *Psychol. Rev.* **74**, 29–39 (1967).

44. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).

45. Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).

46. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* **22**, 55–67 (2021).

47. Jackson, J. C. *et al.* Emotion semantics show both cultural variation and universal structure. *Science* **366**, 1517–1522 (2019).

48. Nook, E. C., Sasse, S. F., Lambert, H. K., McLaughlin, K. A. & Somerville, L. H. Increasing verbal knowledge mediates development of multidimensional emotion representations. *Nat. Hum. Behav.* **1**, 881–889 (2017).

49. Nelson, C. A. & De Haan, M. A neurobehavioral approach to the recognition of facial expressions in infancy. in *The Psychology of Facial Expression* (eds. Russell, J. A. & Fernández-Dols, J. M.) 176–204 (Cambridge University Press, 1997). doi:10.1017/CBO9780511659911.010.
50. Ache, J. M. *et al.* Neural Basis for Looming Size and Velocity Encoding in the Drosophila Giant Fiber Escape Pathway. *Curr. Biol.* **29**, 1073-1081.e4 (2019).
51. Hatsopoulos, N., Gabbiani, F. & Laurent, G. Elementary Computation of Object Approach by a Wide-Field Visual Neuron. *Science* **270**, 1000–1003 (1995).
52. Evans, D. A. *et al.* A synaptic threshold mechanism for computing escape decisions. *Nature* **558**, 590–594 (2018).
53. Jolly, E. & Chang, L. J. The Flatland Fallacy: Moving Beyond Low–Dimensional Thinking. *Top. Cogn. Sci.* **11**, 433–454 (2019).
54. OpenCV. (2022).
55. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. in *Lecture Notes in Computer Science* vol. 2749 363–370 (Springer, Berlin, Heidelberg, 2003).
56. PyTorch. (2022).
57. Schneider, K. A. & Kastner, S. Visual Responses of the Human Superior Colliculus: A High-Resolution Functional Magnetic Resonance Imaging Study. *J. Neurophysiol.* **94**, 2491–2503 (2005).
58. McIlwain, J. T. Distributed spatial coding in the superior colliculus: A review. *Vis. Neurosci.* **6**, 3–13 (1991).
59. FIL Methods Group. SPM12. (2020).
60. MathWorks. MATLAB. (2022).
61. Friston, K. J., Ashburner, J., Frith, C. D. & Frackowiak, R. S. J. Spatial Registration and Normalization of Images. *Hum. Brain Mapp.* **2**, 165–189 (1995).

598 62. Ashburner, J., Neelin, P., Collins, D. L., Evans, A. & Friston, K. Incorporating Prior Knowledge into
599 Image Registration. *NeuroImage* **6**, 344–352 (1997).
600 63. Ashburner, J. & Friston, K. J. Nonlinear spatial normalization using basis functions. *Hum. Brain*
601 *Mapp.* **7**, 254–266 (1999).
602 64. Neuroimaging_Pattern_Masks. (2023).
603 65. Kragel, P. A. *et al.* A human colliculus-pulvinar-amygdala pathway encodes negative emotion.
604 *Neuron* **109**, 2404-2412.e5 (2021).
605 66. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178
606 (2016).
607 67. Rohart, F., Gautier, B., Singh, A. & Cao, K.-A. L. mixOmics: An R package for 'omics feature selection
608 and multiple data integration. *PLOS Comput. Biol.* **13**, e1005752 (2017).
609 68. Kuhn, M. & Wickham, H. Tidymodels: a collection of packages for modeling and machine learning
610 using tidyverse principles. (2020).
611 69. R Core Team. R: A language and environment for statistical computing. (2022).
612 70. Esterman, M., Tamber-Rosenau, B. J., Chiu, Y.-C. & Yantis, S. Avoiding non-independence in fMRI
613 data analysis: Leave one subject out. *NeuroImage* **50**, 572–576 (2010).
614 71. Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E. & Smith, S. M. Multi-level block
615 permutation. *NeuroImage* **123**, 253–268 (2015).
616 72. Adobe Systems. Adobe Photoshop CS5.
617 73. Bacher, L. F. & Smotherman, W. P. Systematic temporal variation in the rate of spontaneous eye
618 blinking in human infants. *Dev. Psychobiol.* **44**, 140–145 (2004).
619 74. Freedman, D. & Lane, D. A Nonstochastic Interpretation of Reported Significance Levels. *J. Bus.*
620 *Econ. Stat.* **1**, 292–298 (1983).

75. Anderson, M. J. & Robinson, J. Permutation Tests for Linear Models. *Aust. N. Z. J. Stat.* **43**, 75–88 (2001).

Additional information

Acknowledgments: We thank Baohua Zhou for assistance with configuring the shallow neural network model, and the ECCO Lab at Emory University for helpful feedback on the project. This work was supported by the National Institutes of Health Institutional Research and Career Development Award (IRACDA) grant K12GM000680 to MKT.

Author contributions:

Conceptualization: PAK, MKT
Methodology: PAK, SFL, MKT
Investigation: VA
Formal analysis: PAK, MKT
Software: MKT
Visualization: MKT
Project administration: PAK, SFL
Supervision: PAK, SFL
Writing – original draft: PAK, MKT
Writing – review & editing: VA, PAK, SFL, MKT

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data and materials that were generated for this study will be posted on Open Science Framework and all code will be posted on GitHub by time of publication.