Affective Science (2023) 4:550–562
https://doi.org/10.1007/s42761-023-00215-z

SHORT REVIEW

Advancing Naturalistic Affective Science with Deep Learning

Chujun Lin[1] · Landry S. Bulls[1] · Lindsey J. Tepfer[1] · Amisha D. Vyas[1] · Mark A. Thornton[1]

Received: 11 January 2023 / Accepted: 3 August 2023 / Published online: 25 August 2023

Abstract
People express their own emotions and perceive others' emotions via a variety of channels, including facial movements, body gestures, vocal prosody, and language. Studying these channels of affective behavior offers insight into both the experience and perception of emotion. Prior research has predominantly focused on studying individual channels of affective behavior in isolation using tightly controlled, non-naturalistic experiments. This approach limits our understanding of emotion in more naturalistic contexts where different channels of information tend to interact. Traditional methods struggle to address this limitation: manually annotating behavior is time-consuming, making it infeasible to do at large scale; manually selecting and manipulating stimuli based on hypotheses may neglect unanticipated features, potentially generating biased conclusions; and common linear modeling approaches cannot fully capture the complex, nonlinear, and interactive nature of real-life affective processes. In this methodology review, we describe how deep learning can be applied to address these challenges to advance a more naturalistic affective science. First, we describe current practices in affective research and explain why existing methods face challenges in revealing a more naturalistic understanding of emotion. Second, we introduce deep learning approaches and explain how they can be applied to tackle three main challenges: quantifying naturalistic behaviors, selecting and manipulating naturalistic stimuli, and modeling naturalistic affective processes. Finally, we describe the limitations of these deep learning methods, and how these limitations might be avoided or mitigated. By detailing the promise and the peril of deep learning, this review aims to pave the way for a more naturalistic affective science.

Keywords Deep learning · Affective science · Person perception · Generalizability · Cognitive modeling

Humans express and recognize emotions using multiple channels in contextually flexible ways (Cowen & Keltner, 2021; Kret et al., 2020; Neal & Chartrand, 2011; Niedenthal et al., 2009; Nummenmaa et al., 2014). These channels include facial movements (Coles et al., 2019; Ekman, 1993; Namba et al., 2022; Wood et al., 2016), body language (C. Ferrari et al., 2022; Poyo Solanas et al., 2020; Reed et al., 2020; Wallbott, 1998), and the tone and content of speech (Bachorowski & Owren, 1995; Beukeboom, 2009; Hawk et al., 2009; Ponsot et al., 2018). Context – both the physical and human environment – also plays a key role (Greenaway et al., 2018; Ngo & Isaacowitz, 2015; Whitesell & Harter, 1996).

Prior research focusing on each individual channel of affective information has advanced a mechanistic understanding of emotion. However, this approach limits generalizability to real-world contexts where different channels of information naturally interact (Yarkoni, 2022). This relatively non-naturalistic tradition of affective research stems in part from technical barriers related to analyzing emotion in more naturalistic contexts. Here, we introduce how deep learning could be applied to overcome these barriers. By understanding its promises and being mindful of its limitations, researchers may use deep learning to advance a more naturalistic affective science.

Current Practices in Affective Research

Before we introduce deep learning applications for affective research, we discuss current practices and challenges. We focus on three topics: quantifying behavior, optimizing stimuli, and modeling affective processes.

Landry S. Bulls, Lindsey J. Tepfer and Amisha D. Vyas contributed equally to this work.

Handling Editor: Michelle (Lani) Shiota

✉ Chujun Lin
Chujun.Lin@Dartmouth.edu

[1] Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA

Researchers commonly quantify behavior through manual annotation. For instance, they annotate the activation of facial muscles (e.g., cheek raiser) on participants' faces using the Facial Action Coding System (Ekman & Friesen, 1978; Girard et al., 2013; Kilbride & Yarczower, 1983). Others manually measure the joint angles of mannequin figures and point-light displays to study body language (Atkinson et al., 2004; Coulson, 2004; Roether et al., 2009; Thoresen et al., 2012).

However, manual annotation is time-consuming (Cohn et al., 2007). This limits the quantity and frequency of annotation. For instance, it would be infeasible to annotate every frame in a large set of videos. As a result, prior research has disproportionately used small samples of static, artificial stimuli (Aviezer et al., 2008, 2012; Benitez-Quiroz et al., 2018; Cowen et al., 2021; McHugh et al., 2010).

Existing research also uses computational tools for quantifying behaviors. For instance, computational models of faces and facial expressions help reveal diagnostic features that people use to infer emotions from faces (Blanz & Vetter, 1999; Jack & Schyns, 2017; Martinez, 2017). Using digital equipment, researchers measure vocal features of speech such as amplitude and frequency (Scherer, 1995, 2003). Models that link these vocal features to emotions further enable researchers to manipulate emotional vocalizations during conversations in real time (Arias et al., 2021). Researchers also investigate the emotional content of speech by computing the frequencies of emotion laden words (e.g., words that commonly express happiness) to perform sentiment analysis (Crossley et al., 2017; Pennebaker & Francis, 1999).

These computational tools work well with highly controlled stimuli (e.g., high quality audiovisual recordings). However, they struggle with naturalistic stimuli, such as real-world conversations where speakers talk over each other in noisy environments. More generally, many of these computational models are based on theory-driven features (e.g., facial/vocal features, or words that researchers think might be associated with emotions), which could miss important emotional features that researchers do not anticipate.

The challenge of quantifying behaviors leads to the further challenge of optimizing naturalistic stimuli that portray these behaviors. Representative sampling is necessary to make inferences from samples to populations. This is widely understood by psychologists, and the field is making increasing efforts to recruit participants from more diverse populations (Barrett, 2020; Henrich et al., 2010; Rad et al., 2018). However, it is less widely appreciated that the need for representative sampling also applies to stimuli (Brunswik, 1955).

The lack of tools for quantifying behavior in naturalistic stimuli (e.g., video recording of participants' emotional responses in conversations) makes it difficult to systematically select and manipulate stimuli. As a result, much prior research relies on manual selection (e.g., selecting recordings based on basic emotions) and manual manipulation (e.g., changing the joint angles of point-light displays based on hypotheses). These methods introduce researchers' preconceived beliefs into experimental designs, and may lead to conclusions that favor those beliefs.

The technical barriers to quantifying behavior and optimizing stimuli contribute to a third challenge, modeling naturalistic affective processes. The mind integrates different channels of affective information in complex and contextual manners. For instance, these integrations may be nonlinear (e.g., when paired with a high-pitched vocalization, both wide and squinting eyes could signal frustration). Different subsets of information streams may be integrated at different stages (e.g., identity and context first, and then with facial expressions). Common linear modeling approaches cannot fully capture these complex processes.

The Promise of Deep Learning for Affective Research

Machine learning is an umbrella term for the practice of training computer algorithms to discover patterns in and make predictions about data (Table 1). Deep learning is a subset of machine learning based on deep neural networks (DNNs) (Rumelhart et al., 1986). DNNs consist of networks of artificial neurons, roughly akin to neurons, and connections, representing synapses. By optimizing the connection (i.e., weights) between neurons the model learns a mapping between the inputs and the outputs that minimizes the prediction errors during the training process (Fig. 1). There are a wide variety of DNN architectures used for solving different computational problems (Table 1).

Here, we describe how DNNs could help address the challenges of behavior quantification, stimuli optimization, and cognitive modeling (Fig. 2). First, many pre-trained DNN models can be readily applied to quantify different channels of affective behavior and do not require any additional model training (Table 2). These models have four distinct advantages over manual annotations and existing computational models.

First, many pre-trained DNNs are efficient to use. For instance, some face annotation DNNs can quantify action units, facial key points, and head poses across thousands of frames of a video in a few minutes (Baltrusaitis et al., 2018; Benitez-Quiroz et al., 2016). This speed advantage creates new possibilities. For instance, using these tools, researchers could predict participants' subjective experience of emotions in real time based on the behavioral quantifications of

Table 1 Comparing deep learning with other machine learning methods

Goal	Examples of Deep Learning Methods	Examples of Other Machine Learning Methods
Regression (Linear)	Single-layer Perceptron with Linear Activations	Linear Regression, Ridge, Lasso
Regression (Nonlinear)	Multilayer Perceptron with Nonlinear Activations	Generalized Linear Model, Polynomial Regression
Regression (Time series, Sequences)	Long Short-term Memory Network, Transformer	Autoregressive models, Hidden Markov Model
Classification	Convolutional Neural Network	Support Vector Machine, Random Forest
Dimension Reduction (Linear)	Autoencoder with Linear Activations	Principal Component Analysis, Exploratory Factor Analysis
Dimension Reduction (Nonlinear)	Autoencoder with Nonlinear Activations, Self-supervised Model	T-distributed Stochastic Neighbor Embedding, Uniform Manifold Approximation and Projection
Clustering	N/A (deep learning can facilitate clustering but does not itself return categorical outputs)	K-Mean, Hierarchical Clustering, Gaussian Mixture Model
Cognitive Models	Spiking Network	Drift Diffusion Model
Agentic Models	Deep reinforcement learning	Reinforcement learning

Examples represent common use cases; they are neither exclusive nor exhaustive

them from video recordings (Li et al., 2021) and use these predictions to time exactly when to introduce experimental manipulations (Fig. 2A).

Second, using pre-trained DNNs reduces costs. For instance, to study body language of real people in social interactions, traditionally researchers need to acquire expensive devices such as motion capture suits or camera systems (Hart et al., 2018; Zane et al., 2019). In comparison, pre-trained DNNs can annotate body poses and joint positions based on ordinary video recordings (Kocabas et al., 2020, 2021; Rempe et al., 2021). These pre-trained and lightweight DNN models provide accessible alternatives to a broader range of researchers.

Third, pre-trained DNNs for behavioral quantification are well suited for complex, real-world contexts. For instance, many pre-trained DNNs are available for naturalistic speech analysis, including the separation of overlapping speech sources, conversion of speech to text, and the quantification of these text in terms of their meaning (Chernykh & Prikhodko, 2018; Lutati et al., 2022a; C. Wang et al., 2022). They offer researchers more powerful tools to investigate how people communicate emotions in real-world conversations and large text corpora across cultures and languages (Ghosal et al., 2019; Poria et al., 2019; Thornton et al., 2022).

Fourth, pre-trained DNNs for behavioral quantification are flexible to use. For instance, there are a range of pre-trained DNNs for quantifying contexts (physical and human environment). At one extreme, researchers can combine multiple models to quantify different elements in the context, such as the interacting partners' behaviors and the objects present (Bhat et al., 2020). At the other extreme, researchers

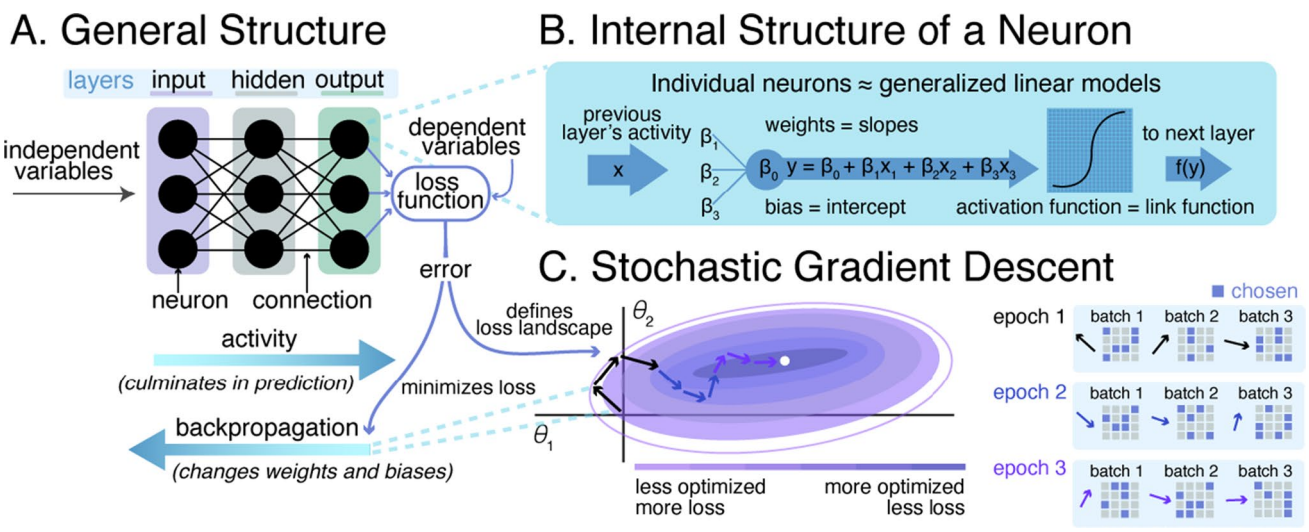

Fig. 1 The structure and training process of a DNN. **A**. The basic components of a DNN. **B**. The computations performed inside a neuron. **C**. The training process for minimizing loss (prediction error) using stochastic gradient descent via backpropagation

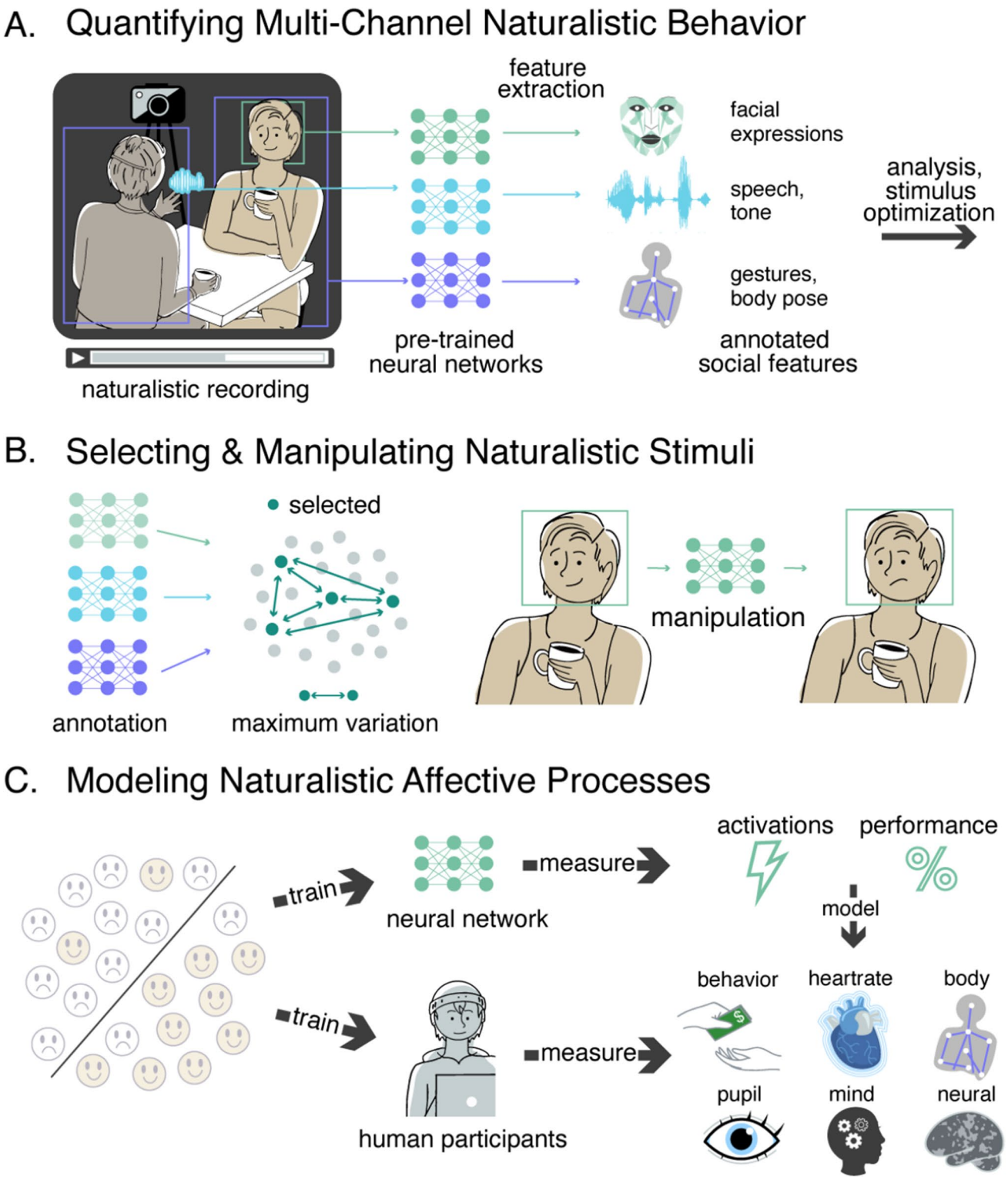

Fig. 2 Applications of DNNs for advancing naturalistic affective research. **A**. DNNs provide a more scalable way to quantify behavior of study participants and stimulus targets in naturalistic contexts. **B**. DNN-based quantifications can support better experimentation by facilitating naturalistic stimulus selection and manipulation. **C**. DNNs are capable of capturing interactive and nonlinear effects, ideal for modeling cognitive/neural mechanisms underlying the subjective experience, physiological responses, and the recognition and expressions of emotions

Table 2 Examples of pre-trained DNN models and DNN architectures for quantifying naturalistic behavior

Modality	Application	Citation
Facial Expression	Emotion Labelling	(Cheong et al., 2021)
		(Wasi et al., 2023)
		(Xue et al., 2022)
		(Wen et al., 2023)
		(Vo et al., 2020)
		(Wang, Peng, et al., 2020)
	Action Unit Annotation	(Cheong et al., 2023)
		(Savchenko, 2022)
		(Luo et al., 2022)
		(Pumarola et al., 2018a)
Body Pose	Pose Estimation	(Kocabas et al., 2021)
		(Xu, Zhang, et al., 2022)
		(Liu et al., 2022)
		(Cai et al., 2020)
		(Sun et al., 2019)
		(Xiao et al., 2018)
Speech	Speech Separation	(Lutati et al., 2023)
		(Lutati et al., 2022b)
		(Subakan et al., 2021)
		(Luo et al., 2020)
		(Luo & Mesgarani, 2019)
	Speech-to-Text	(Bain et al., 2023)
		(Gállego et al., 2021)
		(Conneau et al., 2021)
		(Wang, Tang, et al., 2020)
		(Park et al., 2019)
Text	Text Embedding	(He et al., 2021)
		(Gao et al., 2021)
		(Lan et al., 2020)
		(Reimers & Gurevych, 2019)
		(Cer et al., 2018)
	Sentiment Analysis	(Heinsen, 2022)
		(Sun et al., 2020)
		(Heinsen, 2020)
		(Peters et al., 2018)

can extract a global description of the scene (Krishna et al., 2017). The ability to quantify context in these different operational levels makes it possible to study the effects of context on emotion recognition, emotion expression, and the subjective, physiological, and neural components of emotion more precisely.

Applying deep learning to quantify behavior benefits stimulus optimization efforts as well (Fig. 2B). Imagine a case in which researchers are investigating how storytelling evokes emotional experiences. Selecting storytelling videos that are representative of the diverse storytelling that people encounter in daily life will facilitate a more generalizable conclusion (Fig. 2B, left). The multi-channel behavioral quantifications from DNNs can help achieve this goal. Specifically, researchers could first scrape a large number of real-world storytelling videos from the internet; then quantify multi-channel information in each video (e.g., face, body, speech, context) using deep learning models; and finally, apply the maximum variation sampling procedure (Patton, 1990) to select a subset of stimuli from every part of the psychological space.

To better understand the causal relation between different channels of information and affective responses, researchers may wish to manipulate stimuli beyond selecting them (Fig. 2B right). Deep learning models can also manipulate naturalistic stimuli realistically and in real time (Xu, Hong, et al., 2022). For instance, researchers could manipulate the facial expressions of participants as they spoke to each other over a video call and measure how one participant's manipulated facial expressions influence the other partner's subjective experience of emotions or physiological responses. This would allow for controlled experiments on naturalistic conversations through the medium of the conversation itself, rather than imposing an external intrusion upon it (e.g., prompts to change conversation topics).

Researchers can also use deep learning to achieve experimental control over naturalistic stimuli by synthesizing novel stimuli that never existed in the real world (Balakrishnan et al., 2018; Daube et al., 2021; Guo et al., 2023; Liu et al., 2021; Masood et al., 2023; Pumarola et al., 2018b; Ren & Wang, 2022; Roebel & Bous, 2022; Schyns et al., 2023; Wang et al., 2018; Yu et al., 2019). These tools can generate high-quality, realistic images, audios, and videos of any combination of features that the researchers might be interested in, some even in real time. This can provide an unprecedented level of control to researchers while still retaining naturalism in the stimuli.

Finally, deep learning could advance a computational understanding of naturalistic affective processes in the mind and brain (Fig. 2C). Many researchers have already applied deep learning to cognitive modeling, such as how information is represented in the visual cortex (Cichy & Kaiser, 2019; Dobs et al., 2022; Khaligh-Razavi & Kriegeskorte, 2014; Kohoutová et al., 2020; Konkle & Alvarez, 2022; Mehta et al., 2020; Perconti & Plebe, 2020; Richards et al., 2019; Saxe et al., 2021; Su et al., 2020).

Research has started applying deep learning to model affective cognition (Kragel et al., 2019; Thornton et al., 2023). Three qualities of DNNs make them a promising avenue for advancing a naturalistic understanding of affective processes. First, by virtue of their nonlinear activation functions and multi-layered structure (Fig. 1), DNN models excel

at discovering complex interactions among both observable variables (e.g., affective behaviors) and latent variables. Given the importance of latent variables (e.g., emotions) and the complex interactions between behaviors and contexts, this feature is essential for building realistic cognitive models of affective processing.

Second, DNN models can predict multidimensional dependent variables in a single integrated model. Unlike common regression-based models, which typically have scalar outputs, the dependent variables in DNN models can be scalars, vectors, or multidimensional arrays. Moreover, DNN models can capitalize on the structure of the data, modeling both spatial relationships (e.g., via convolutions) and temporal relationships (e.g., via recurrence). Although one can find these individual pieces in other bespoke statistical models (Table 1), arguably nothing rivals the flexibility of deep learning at combining them into a single computationally efficient package. Since both the inputs (others' naturalistic behavior) and outputs of affective processes (subjective, physiological, and neural components of emotions) are frequently multidimensional and complexly structured in time and space, this flexibility makes deep learning useful for affective modeling.

Third, DNN models provide a useful framework for simulating causal effects. For instance, to understand how different types of affective behaviors (e.g., face and voice) interact to express emotions, one can manipulate a DNN's architecture so that different cues are allowed to interact in different layers of the models. These manipulations are impossible to do in the human brain as one cannot simply rewire it at will. Deep learning can also be embedded within embodied agents (Arulkumaran et al., 2017) so that researchers can use them to study how affective processes shape action and decision-making as agents learn to causally manipulate their environment.

Limitations of Deep Learning for Affective Research

Despite its promise, deep learning is not a magic box. Understanding the limitations of DNNs will help affective scientists use them effectively. Here, we describe the limitations of DNNs for behavior quantification, stimuli optimization, and cognitive modeling.

First, although the accuracy of pre-trained DNNs for annotating affective behavior is relatively high, many of them have yet to achieve human-level accuracy. For instance, the accuracy of estimating three-dimensional facial expressions and body movements is constrained by the two-dimensional inputs that these models are trained on (images and videos). However, given the fast pace of deep learning improvements (Fig. 3), there is reason to be optimistic about improvement in this regard.

Second, the benefits of using pre-trained DNNs for behavioral annotation vary with the context. For instance, many of these models annotate only a subset of behavioral features that researchers might be interested in, such as only 20 out of 46 facial action units. The performance of these models may be significantly reduced in certain situations. For instance, DNN audio source separation may fail when the quality of audio recordings is low. These models also struggle to generalize. For instance, a facial expression classification model that performs well in the conditions it was trained on (e.g., frontal, well-lit, adult faces), may perform poorly when applied to different conditions (Cohn et al., 2019). Careful accuracy and bias auditing should be part of any study relying on deep learning as an objective quantification tool.

Third, DNNs are susceptible to social biases. These biases may result from the composition of the training dataset (e.g., having more samples for certain ethnicities than others), the bias of the humans who provided the training labels (e.g., stereotyped associations), and/or the architecture of the algorithm itself (Mehrabi et al., 2021; Shankar et al., 2017). For instance, some algorithms wrongly assign more negative emotions to Black men's faces than White men's faces, reflecting a stereotype widespread in the US (Kim et al., 2021; Rhue, 2018; Schmitz et al., 2022).

Since the application of stimulus optimization uses outputs from pre-trained DNNs, the above limitations of the quantification models can carry through to influence stimulus selection. For instance, if a voice quantification model has been overtrained on male versus female speech, it may represent male voices as more distinct from each other than female voices. Applying maximum variation sampling based on these quantifications might thus lead to over-sampling of male speech.

Maximum variation sampling is also susceptible to class imbalance (Van Calster et al., 2019). For instance, if the initial stimulus set has significantly more positive valence stories than negative ones, then more positive stories will be selected more frequently. However, both issues with stimulus selection can be mitigated with stratified maximum variation sampling (e.g., applying the procedure to male and female speech separately, or positive and negative stories separately) (Lin et al., 2021, 2022; Lin & Thornton, 2023).

While some applications of deep learning are approaching maturity (i.e., achieving high-level accuracy), such as behavior quantification (Fig. 3), others are just emerging, such as manipulating and synthesizing realistic stimuli in real time. At present, applying DNNs for manipulating and synthesizing stimuli are most reliable for features that the models have been exposed to (Hosseini et al., 2017; Papernot et al., 2016). For instance, if an algorithm has never been trained on the motion of jumping, its synthesis of people jumping with joy is unlikely to look realistic.

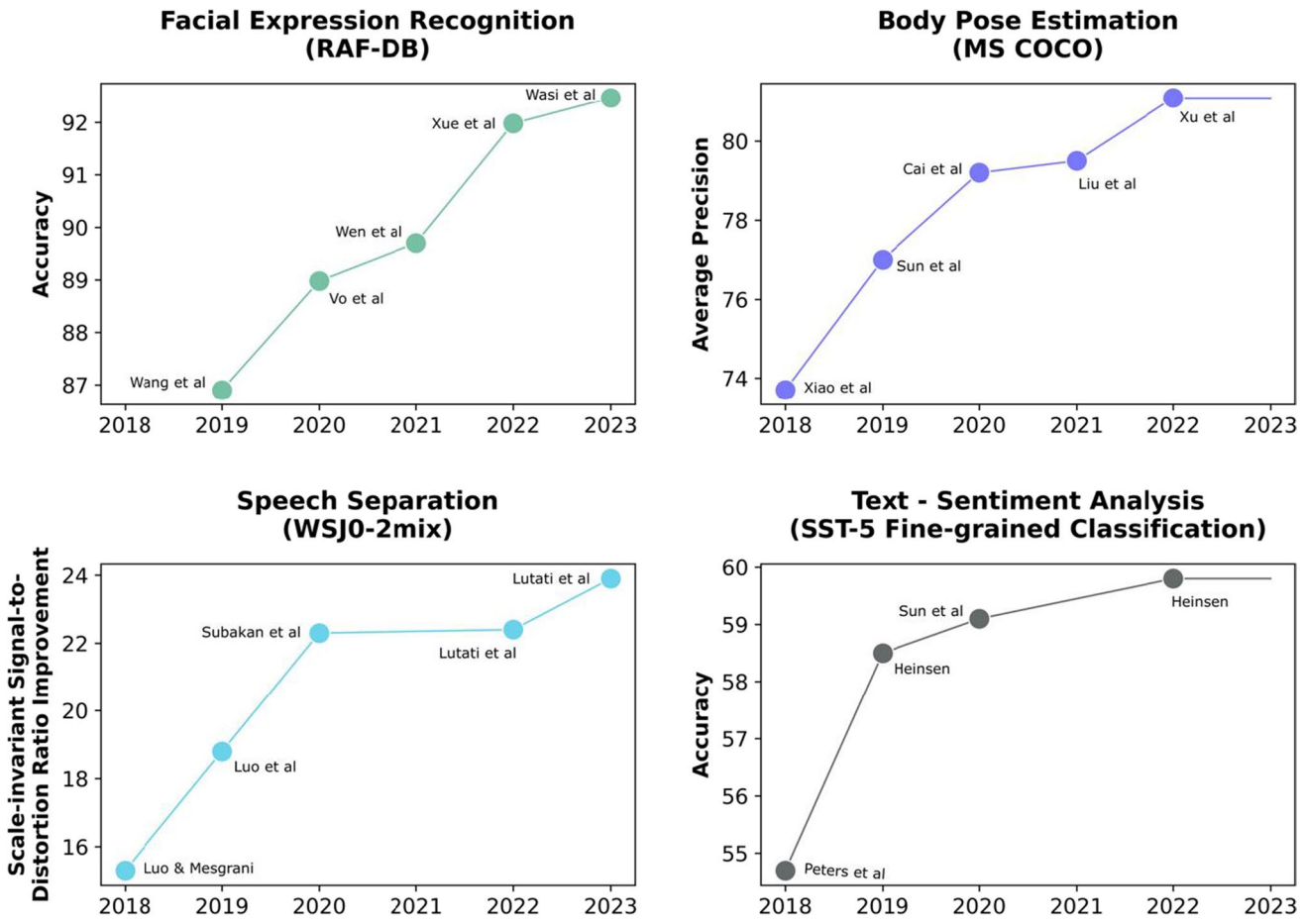

Fig. 3 Improvements of DNNs for behavioral quantification over time. Title indicates the behavior channel and the corresponding benchmark dataset that the models were evaluated on. X-axis indicates the year the models were published. Y-axis indicates the metric for measuring model performance. Data reflect benchmarks reported on paperswithcode.com (Papers with Code, n.d.)

When applying DNNs for affective modeling, in addition to the overall level of accuracy, researchers should carefully consider the types of errors that DNNs make, which may be systematically different from those of humans. Researchers should also caution equating deep learning models to the human mind and brain. Correlations between human performance and DNNs do not indicate that the two systems share similar causal mechanisms (Bowers et al., 2022; Schyns et al., 2022).

Finally, researchers should also distinguish between the inherent versus current limitations of DNNs. For instance, many existing DNN models are trained on aggregate-level data and thus cannot represent individual differences in affective processes. However, with the proper inputs (e.g., individual-level perceptions with individual difference measures), DNNs could in principle model individual differences in affective processes.

Besides the limitations highlighted for each of the three applications above, we have summarized the most prominent limitations of deep learning in general, alongside potential mitigation strategies for each of them (Table 3).

Conclusion

In this review, we have provided a brief introduction to how deep learning could be applied to tackle challenges in affective science. We focused on three main applications: behavior quantification, stimuli optimization, and affective modeling. These applications can advance naturalistic research on the verbal and nonverbal expressions of emotions, the recognition of emotions, and the subjective, physiological, and neural components of affective experiences. We encourage interested readers to explore other works that provide detailed primers on how to use these tools to their fullest (Pang et al., 2020; Thomas et al., 2022; Urban & Gates, 2021; Yang & Molano-Mazón, 2021). With deep learning tools in hand, they will stand

Table 3 Limitations of deep learning

Limitation	Explanation	Mitigation Strategies
Social bias	Worse, or systematically different, performance for marginalized groups; Reflects bias in dataset composition, annotation, or algorithm construction	Perform bias audits; Retrain model with less problematic data or algorithm; Critically consider goals and (mis)uses of algorithms
Causal Inference	Model performance may reflect causal or confounding relationships in data, and model cannot distinguish them	Continue using normal causal identification strategies (e.g., experiments, instruments)
Interpretability	Large number of parameters and nonlinear relationships render models opaque/inexplainable in human terms	Visualize units' "receptive fields"; "lesion" parts of model or augment data to reveal function
Costly to train	Large, high-quality training datasets can be expensive to collect/create; Training large models can require expensive hardware and incur large electricity costs	Use pretrained models; Use smaller "distilled" models that offer similar performance with fewer parameters; Share costs with other researchers
Performance	Most existing models still perform worse than human gold standard; The types of errors made by models may be very different from those made by humans	Wait for state-of-art to improve; tolerate scale vs. accuracy tradeoff; examine error patterns
Generalization	Model performance generally degrades under "distribution shift" – i.e., models can interpolate within the examples they have been trained on, but often fail to extrapolate to new regions of the feature/task space; Versions of the same model trained on the same data with different random seeds can generalize very differently	Audit performance on own data; Fine-tune pretrained models to improve generalization to specific use case; Avoid deploying models to cases far beyond their training set; Stress test different versions of the same model
Symbolic Reasoning	Models cannot generically solve non-differentiable or symbolic problems, and unsupervised clustering; Large models can memorize specific symbol patterns but cannot generalize rules	Use symbolic AI; Use hybrid deep learning-symbolic AI systems; Avoid non-differentiable problems; Audit for memorization
Feedback	Models are feed-forward only, meaning they cannot model feedback processes that occur in the brain; Limits ability to model temporal dynamics	Use non-feed-forward ANNs (e.g., spiking networks); Model longer timescales (e.g., time course of learning)
Technical skills	Relatively high level of programming proficiency; acquisition of many skills specific to deep learning	Create and use open learning resources (e.g., Jupyter Books); Amend graduate curriculum

poised to substantially expand our understanding of emotion in more naturalistic contexts.

Additional Information

Funding Not applicable

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Availability of data and material Not applicable

Code availability Not applicable

Authors' contributions Not applicable

Ethics approval Not applicable

Consent to participate Not applicable

Consent for publication Not applicable

References

Arias, P., Rachman, L., Liuni, M., & Aucouturier, J.-J. (2021). Beyond correlation: Acoustic transformation methods for the experimental study of emotional voice and speech. *Emotion Review, 13*(1), 12–24. https://doi.org/10.1177/1754073920934544

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine, 34*(6), 26–38. https://doi.org/10.1109/MSP.2017.2743240

Atkinson, A. P., Dittrich, W. H., Gemmell, A. J., & Young, A. W. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception, 33*(6), 717–746. https://doi.org/10.1068/p5096

Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., & Bentin, S. (2008). Angry, disgusted, or afraid?: Studies on the malleability of emotion perception. *Psychological Science, 19*(7), 724–732. https://doi.org/10.1111/j.1467-9280.2008.02148.x

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science, 338*(6111), 1225–1229. https://doi.org/10.1126/science.1224313

Bachorowski, J.-A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science, 6*(4), 219–224. https://doi.org/10.1111/j.1467-9280.1995.tb00596.x

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). *WhisperX: Time-accurate speech transcription of long-form audio* (arXiv:2303.00747). arXiv. https://doi.org/10.48550/arXiv.2303.00747

Balakrishnan, G., Zhao, A., Dalca, A. V., Durand, F., & Guttag, J. (2018). *Synthesizing images of humans in unseen poses* (arXiv:1804.07739; version 1). arXiv. http://arxiv.org/abs/1804.07739

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). OpenFace 2.0: Facial behavior analysis toolkit. *2018 13th IEEE international conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66. https://doi.org/10.1109/FG.2018.00019

Barrett, H. C. (2020). Towards a cognitive science of the human: Cross-cultural approaches and their urgency. *Trends in Cognitive Sciences, 24*(8), 620–638. https://doi.org/10.1016/j.tics.2020.05.007

Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2016). EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5562–5570. https://doi.org/10.1109/CVPR.2016.600

Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2018). Facial color is an efficient mechanism to visually transmit emotion. *Proceedings of the National Academy of Sciences, 115*(14), 3581–3586. https://doi.org/10.1073/pnas.1716084115

Beukeboom, C. J. (2009). When words feel right: How affective expressions of listeners change a speaker's language use. *European Journal of Social Psychology, 39*(5), 747–756. https://doi.org/10.1002/ejsp.572

Bhat, G., Danelljan, M., Van Gool, L., & Timofte, R. (2020). Know your surroundings: Exploiting scene information for object tracking. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer vision – ECCV 2020 (Vol. 12368, pp. 205–221)*. Springer International Publishing. https://doi.org/10.1007/978-3-030-58592-1_13

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '99*, 187–194. https://doi.org/10.1145/311535.311556

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 1–74. https://doi.org/10.1017/S0140525X22002813

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*, 193–217. https://doi.org/10.1037/h0047470

Cai, Y., Wang, Z., Luo, Z., Yin, B., Du, A., Wang, H., Zhang, X., Zhou, X., Zhou, E., & Sun, J. (2020). Learning delicate local representations for multi-person pose estimation. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer vision – ECCV 2020 (Vol. 12348, pp. 455–472)*. Springer International Publishing. https://doi.org/10.1007/978-3-030-58580-8_27

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). *Universal sentence encoder* (arXiv:1803.11175). arXiv. https://doi.org/10.48550/arXiv.1803.11175

Cheong, J. H., Jolly, E., Xie, T., Byrne, S., Kenney, M., & Chang, L. J. (2021). *Py-feat: Python facial expression analysis toolbox* (arXiv:2104.03509). arXiv. http://arxiv.org/abs/2104.03509

Cheong, J. H., Jolly, E., Xie, T., Byrne, S., Kenney, M., & Chang, L. J. (2023). *Py-feat: Python facial expression analysis toolbox* (arXiv:2104.03509). arXiv. https://doi.org/10.48550/arXiv.2104.03509

Chernykh, V., & Prikhodko, P. (2018). *Emotion recognition from speech with recurrent neural networks* (arXiv:1701.08071; version 2). arXiv. http://arxiv.org/abs/1701.08071

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences, 23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Cohn, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the facial action coding system. In *Handbook of emotion elicitation and assessment* (pp. 203–221). Oxford University Press.

Cohn, J. F., Ertugrul, I. O., Chu, W. S., Girard, J. M., & Hammal, Z. (2019). Affective facial computing: Generalizability across domains. In *Multimodal behavior analysis in the wild: Advances and challenges* (pp. 407–441). Academic Press.

Coles, N. A., Larsen, J. T., & Lench, H. C. (2019). A meta-analysis of the facial feedback literature: Effects of facial feedback on

emotional experience are small and variable. *Psychological Bulletin, 145*(6), 610–651. https://doi.org/10.1037/bul0000194

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. *Interspeech 2021*, 2426–2430. https://doi.org/10.21437/Interspeech.2021-329

Coulson, M. (2004). Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior, 28*, 117–139. https://doi.org/10.1023/B:JONB.0000023655.25550.be

Cowen, A. S., & Keltner, D. (2021). Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences, 25*(2), 124–136. https://doi.org/10.1016/j.tics.2020.11.004

Cowen, A. S., Keltner, D., Schroff, F., Jou, B., Adam, H., & Prasad, G. (2021). Sixteen facial expressions occur in similar contexts worldwide. *Nature, 589*(7841), 7841. https://doi.org/10.1038/s41586-020-3037-7

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods, 49*(3), 803–821. https://doi.org/10.3758/s13428-016-0743-z

Daube, C., Xu, T., Zhan, J., Webb, A., Ince, R. A. A., Garrod, O. G. B., & Schyns, P. G. (2021). Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns, 2*(10), 100348. https://doi.org/10.1016/j.patter.2021.100348

Dobs, K., Martinez, J., Kell, A. J. E., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances, 8*(11), eabl8913. https://doi.org/10.1126/sciadv.abl8913

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system*. Environmental Psychology & Nonverbal Behavior.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(3), 376–379.

Ferrari, C., Ciricugno, A., Urgesi, C., & Cattaneo, Z. (2022). Cerebellar contribution to emotional body language perception: A TMS study. *Social Cognitive and Affective Neuroscience, 17*(1), 81–90. https://doi.org/10.1093/scan/nsz074

Gállego, G. I., Tsiamas, I., Escolano, C., Fonollosa, J. A. R., & Costa-jussà, M. R. (2021). End-to-end speech translation with pre-trained models and adapters: UPC at IWSLT 2021. *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, 110–119. https://doi.org/10.18653/v1/2021.iwslt-1.11

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence Embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. (2019). *DialogueGCN: A graph convolutional neural network for emotion recognition in conversation* (arXiv:1908.11540). arXiv. http://arxiv.org/abs/1908.11540

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S., & Rosenwald, D. P. (2013). Social risk and depression: Evidence from manual and automatic facial expression analysis. *Proceedings of the ... International Conference on Automatic Face and Gesture Recognition*, 1–8. https://doi.org/10.1109/FG.2013.6553748

Greenaway, K. H., Kalokerinos, E. K., & Williams, L. A. (2018). Context is everything (in emotion research). *Social and Personality Psychology Compass, 12*(6), e12393. https://doi.org/10.1111/spc3.12393

Guo, Z., Leng, Y., Wu, Y., Zhao, S., & Tan, X. (2023). Prompttts: Controllable text-to-speech with text descriptions. *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096285

Hart, T. B., Struiksma, M. E., van Boxtel, A., & van Berkum, J. J. A. (2018). Emotion in stories: Facial EMG evidence for both mental simulation and moral evaluation. F*rontiers in Psychology, 9*, 613. https://doi.org/10.3389/fpsyg.2018.00613

Hawk, S. T., van Kleef, G. A., Fischer, A. H., & van der Schalk, J. (2009). "Worth a thousand words": Absolute and relative decoding of nonlinguistic affect vocalizations. *Emotion, 9*, 293–305. https://doi.org/10.1037/a0015178

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with disentangled attention* (arXiv:2006.03654). arXiv. https://doi.org/10.48550/arXiv.2006.03654

Heinsen, F. A. (2020). *An algorithm for routing capsules in all domains* (arXiv:1911.00792). arXiv. https://doi.org/10.48550/arXiv.1911.00792

Heinsen, F. A. (2022). *An algorithm for routing vectors in sequences* (arXiv:2211.11754). arXiv. https://doi.org/10.48550/arXiv.2211.11754

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hosseini, H., Xiao, B., Jaiswal, M., & Poovendran, R. (2017). On the limitation of convolutional neural networks in recognizing negative images. *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, 352–358. https://doi.org/10.1109/ICMLA.2017.0-136

Jack, R. E., & Schyns, P. G. (2017). Toward a social psychophysics of face communication. *Annual Review of Psychology, 68*(1), 269–297. https://doi.org/10.1146/annurev-psych-010416-044242

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kilbride, J. E., & Yarczower, M. (1983). Ethnic bias in the recognition of facial expressions. *Journal of Nonverbal Behavior, 8*(1), 27–41. https://doi.org/10.1007/BF00986328

Kim, E., Bryant, D., Srikanth, D., & Howard, A. (2021). Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 638–644. https://doi.org/10.1145/3461702.3462609

Kocabas, M., Athanasiou, N., & Black, M. J. (2020). *VIBE: Video inference for human body pose and shape estimation* (arXiv:1912.05656). arXiv. https://doi.org/10.48550/arXiv.1912.05656

Kocabas, M., Huang, C.-H. P., Hilliges, O., & Black, M. J. (2021). PARE: Part attention Regressor for 3D human body estimation. *IEEE/CVF International Conference on Computer Vision (ICCV),* 11107–11117. https://doi.org/10.1109/ICCV48922.2021.01094

Kohoutová, L., Heo, J., Cha, S., Lee, S., Moon, T., Wager, T. D., & Woo, C.-W. (2020). Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nature Protocols, 15*(4), Article 4. https://doi.org/10.1038/s41596-019-0289-5

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications, 13*(1), 491. https://doi.org/10.1038/s41467-022-28091-4

Kragel, P. A., Reddan, M. C., LaBar, K. S., & Wager, T. D. (2019). Emotion schemas are embedded in the human visual system. *Science Advances, 5*(7):eaaw4358

Kret, M. E., Prochazkova, E., Sterck, E. H. M., & Clay, Z. (2020). Emotional expressions in human and non-human great apes. *Neuroscience & Biobehavioral Reviews, 115*, 378–395. https://doi.org/10.1016/j.neubiorev.2020.01.027

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017). Dense-captioning events in videos. *IEEE International Conference on Computer Vision (ICCV),* 706–715. https://doi.org/10.1109/ICCV.2017.83

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *Albert: A lite bert. For Self-supervised learning of language representations*.

Li, Q., Liu, Y. Q., Peng, Y. Q., Liu, C., Shi, J., Yan, F., & Zhang, Q. (2021). Real-time facial emotion recognition using lightweight convolution neural network. *Journal of Physics: Conference Series, 1827*(1), 012130. https://doi.org/10.1088/1742-6596/1827/1/012130

Lin, C., Keles, U., & Adolphs, R. (2021). Four dimensions characterize attributions from faces using a representative set of English trait words. *Nature Communications, 12*(1), 1. https://doi.org/10.1038/s41467-021-25500-y

Lin, C., Keles, U., Thornton, M. A., & Adolphs, R. (2022). How trait impressions of faces shape subsequent mental state inferences [registered report stage 1 protocol]. *Nature Human Behaviour* https://doi.org/10.6084/m9.figshare.19664316.v1

Lin, C., & Thornton, M. A. (2023). Evidence for bidirectional causation between trait and mental state inferences. *Journal of Experimental Social Psychology, 108*, 104495. https://doi.org/10.31234/osf.io/ysn3w

Liu, H., Liu, F., Fan, X., & Huang, D. (2022). Polarized self-attention: Towards high-quality pixel-wise mapping. *Neurocomputing, 506*, 158–167. https://doi.org/10.1016/j.neucom.2022.07.054

Liu, M.-Y., Huang, X., Yu, J., Wang, T.-C., & Mallya, A. (2021). Generative adversarial networks for image and video synthesis: Algorithms and applications. *Proceedings of the IEEE, 109*(5), 839–862. https://doi.org/10.1109/JPROC.2021.3049196

Luo, C., Song, S., Xie, W., Shen, L., & Gunes, H. (2022). Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition. *Proceedings of the Thirty-first International Joint Conference on Artificial Intelligence*, 1239–1246. https://doi.org/10.24963/ijcai.2022/173

Luo, Y., Chen, Z., & Yoshioka, T. (2020). Dual-path RNN: Efficient long sequence modeling for time-domain Single-Channel speech separation. *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 46–50. https://doi.org/10.1109/ICASSP40776.2020.9054266

Luo, Y., & Mesgarani, N. (2019). Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27*(8), 1256–1266. https://doi.org/10.1109/TASLP.2019.2915167

Lutati, S., Nachmani, E., & Wolf, L. (2022a). *SepIt: Approaching a Single Channel speech separation bound* (arXiv:2205.11801; version 3). arXiv. http://arxiv.org/abs/2205.11801

Lutati, S., Nachmani, E., & Wolf, L. (2022b). SepIt: Approaching a Single Channel speech separation bound. *Interspeech* 2022, 5323–5327. https://doi.org/10.21437/Interspeech.2022-149

Lutati, S., Nachmani, E., & Wolf, L. (2023). *Separate and diffuse: Using a Pretrained diffusion model for improving source separation* (arXiv:2301.10752). arXiv. https://doi.org/10.48550/arXiv.2301.10752

Martinez, A. M. (2017). Computational models of face perception. *Current Directions in Psychological Science, 26*(3), 263–269. https://doi.org/10.1177/0963721417698535

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence, 53*(4), 3974–4026. https://doi.org/10.1007/s10489-022-03766-z

McHugh, J. E., McDonnell, R., O'Sullivan, C., & Newell, F. N. (2010). Perceiving emotion in crowds: The role of dynamic body postures on the perception of emotion in crowded scenes. *Experimental Brain Research, 204*(3), 361–372. https://doi.org/10.1007/s00221-009-2037-5

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on Bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Mehta, Y., Majumder, N., Gelbukh, A., & Cambria, E. (2020). Recent trends in deep learning based personality detection. *Artificial Intelligence Review, 53*(4), 2313–2339. https://doi.org/10.1007/s10462-019-09770-z

Namba, S., Sato, W., Nakamura, K., & Watanabe, K. (2022). Computational process of sharing emotion: An authentic information perspective. *Frontiers in Psychology, 13*, 849499. https://doi.org/10.3389/fpsyg.2022.849499

Neal, D. T., & Chartrand, T. L. (2011). Embodied emotion perception: Amplifying and dampening facial feedback modulates emotion perception accuracy. *Social Psychological and Personality Science, 2*(6), 673–678. https://doi.org/10.1177/1948550611406138

Ngo, N., & Isaacowitz, D. M. (2015). Use of context in emotion perception: The role of top-down control, cue type, and perceiver's age. *Emotion, 15*(3), 292–302. https://doi.org/10.1037/emo0000062

Niedenthal, P. M., Winkielman, P., Mondillon, L., & Vermeulen, N. (2009). Embodiment of emotion concepts. *Journal of Personality and Social Psychology, 96*(6), 1120–1136. https://doi.org/10.1037/a0015574

Nummenmaa, L., Glerean, E., Hari, R., & Hietanen, J. K. (2014). Bodily maps of emotions. *Proceedings of the National Academy of Sciences, 111*(2), 646–651. https://doi.org/10.1073/pnas.1321664111

Pang, B., Nijkamp, E., & Wu, Y. N. (2020). Deep learning with TensorFlow: A review. *Journal of Educational and Behavioral Statistics, 45*(2), 227–248. https://doi.org/10.3102/1076998619872761

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 372–387. https://doi.org/10.1109/EuroSP.2016.36

Papers with Code. (n.d.). Retrieved June 26, 2023, from https://paperswithcode.com/

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*, 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

Patton, M. Q. (1990). *Qualitative evaluation and research methods, 2nd ed (p. 532)*. Sage Publications, Inc.

Pennebaker, J., & Francis, M. (1999). *Linguistic inquiry and word count*. Lawrence Erlbaum Associates, Incorporated.

Perconti, P., & Plebe, A. (2020). Deep learning and cognitive science. *Cognition, 203*, 104365. https://doi.org/10.1016/j.cognition.2020.104365

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J.-J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences, 115*(15), 3972–3977. https://doi.org/10.1073/pnas.1716090115

Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access, 7*, 100943–100953. https://doi.org/10.1109/ACCESS.2019.2929050

Poyo Solanas, M., Vaessen, M. J., & de Gelder, B. (2020). The role of computational and subjective features in emotional body expressions. *Scientific Reports, 10*(1), 1. https://doi.org/10.1038/s41598-020-63125-1

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. (2018a). GANimation: Anatomically-aware facial animation from a single image. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision – ECCV 2018* (Vol. 11214, pp. 835–851). Springer International Publishing. https://doi.org/10.1007/978-3-030-01249-6_50

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. (2018b). GANimation: Anatomically-aware facial animation from a single image. *European Conference on Computer Vision*, 818–833. https://openaccess.thecvf.com/content_ECCV_2018/html/Albert_Pumarola_Anatomically_Coherent_Facial_ECCV_2018_paper.html

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences, 115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115

Reed, C. L., Moody, E. J., Mgrublian, K., Assaad, S., Schey, A., & McIntosh, D. N. (2020). Body matters in emotion: Restricted body movement and posture affect expression and recognition of status-related emotions. *Frontiers in Psychology*, 11. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01961

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., & Guibas, L. J. (2021). *HuMoR: 3D human motion model for robust pose estimation* (arXiv:2105.04668). arXiv. https://doi.org/10.48550/arXiv.2105.04668

Ren, X., & Wang, X. (2022). *Look outside the room: Synthesizing a consistent long-term 3D scene video from a single image* (arXiv:2203.09457; version 1). arXiv. http://arxiv.org/abs/2203.09457

Rhue, L. (2018). *Racial influence on automated perceptions of emotions* (SSRN scholarly paper 3281765). https://doi.org/10.2139/ssrn.3281765

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience, 22*(11), 11. https://doi.org/10.1038/s41593-019-0520-2

Roebel, A., & Bous, F. (2022). Neural Vocoding for singing and speaking voices with the multi-band excited WaveNet. *Information, 13*(3), 3. https://doi.org/10.3390/info13030103

Roether, C., Omlor, L., Christensen, A., & Giese, M. (2009). Critical features for the perception of emotion from gait. *Journal of Vision, 9*(15) https://jov.arvojournals.org/article.aspx?articleid=2204009

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 6088. https://doi.org/10.1038/323533a0

Savchenko, A. V. (2022). *Frame-level prediction of facial expressions, valence, arousal and action units for Mobile devices* (arXiv:2203.13436). arXiv. https://doi.org/10.48550/arXiv.2203.13436

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience, 22*(1), 1. https://doi.org/10.1038/s41583-020-00395-8

Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice, 9*(3), 235–248.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

Schmitz, M., Ahmed, R., & Cao, J. (2022). *Bias and fairness on multimodal emotion detection algorithms.*

Schyns, P. G., Snoek, L., & Daube, C. (2022). Degrees of algorithmic equivalence between the brain and its DNN models. *Trends in Cognitive Sciences, 26*(12), 1090–1102. https://doi.org/10.1016/j.tics.2022.09.003

Schyns, P. G., Snoek, L., & Daube, C. (2023). Stimulus models test hypotheses in brains and DNNs. *Trends in Cognitive Sciences, 27*(3), 216–217. https://doi.org/10.1016/j.tics.2022.12.003

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World* (arXiv:1711.08536). arXiv. https://doi.org/10.48550/arXiv.1711.08536

Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry, 10*(1), 1. https://doi.org/10.1038/s41398-020-0780-3

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021). Attention is all you need in speech separation. *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 21–25. https://doi.org/10.1109/ICASSP39728.2021.9413901

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5686–5696. https://doi.org/10.1109/CVPR.2019.00584

Sun, Z., Fan, C., Han, Q., Sun, X., Meng, Y., Wu, F., & Li, J. (2020). *Self-explaining structures improve NLP models* (arXiv:2012.01786). arXiv. https://doi.org/10.48550/arXiv.2012.01786

Thomas, A. W., Ré, C., & Poldrack, R. A. (2022). Interpreting mental state decoding with deep learning models. *Trends in Cognitive Sciences, 26*(11), 972–986. https://doi.org/10.1016/j.tics.2022.07.003

Thoresen, J. C., Vuong, Q. C., & Atkinson, A. P. (2012). First impressions: Gait cues drive reliable trait judgements. *Cognition, 124*(3), 261–271. https://doi.org/10.1016/j.cognition.2012.05.018

Thornton, M. A., Rmus, M., Vyas, A. D., & Tamir, D. I. (2023). Transition dynamics shape mental state concepts. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0001405

Thornton, M. A., Wolf, S., Reilly, B. J., Slingerland, E. G., & Tamir, D. I. (2022). The 3d mind model characterizes how people understand mental states across modern and historical cultures. *Affective* Science, 3(1). https://doi.org/10.1007/s42761-021-00089-z

Urban, C., & Gates, K. (2021). Deep learning: A primer for psychologists. *Psychological Methods*, 26(6). https://doi.org/10.1037/met0000374

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., Bossuyt, P., Collins, G. S., Macaskill, P., McLernon, D. J., Moons, K. G. M., Steyerberg, E. W., Van Calster, B., van Smeden, M., Vickers, A. J., & On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Vo, T.-H., Lee, G.-S., Yang, H.-J., & Kim, S.-H. (2020). Pyramid with super resolution for in-the-wild facial expression recognition.

IEEE Access, 8, 131988–132001. https://doi.org/10.1109/ACCESS.2020.3010018

Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology, 28*(6), 879–896.

Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., & Pino, J. (2020). Fairseq S2T: Fast speech-to-text modeling with Fairseq. *Proceedings of the 1st conference of the Asia-Pacific chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, 33–39. https://aclanthology.org/2020.aacl-demo.6

Wang, C., Tang, Y., Ma, X., Wu, A., Popuri, S., Okhonko, D., & Pino, J. (2022). *Fairseq S2T: Fast speech-to-text modeling with fairseq* (arXiv:2010.05171; version 2). arXiv. http://arxiv.org/abs/2010.05171

Wang, K., Peng, X., Yang, J., Meng, D., & Qiao, Y. (2020). Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing, 29*, 4057–4069. https://doi.org/10.1109/TIP.2019.2956143

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., & Catanzaro, B. (2018). *Video-to-video synthesis* (arXiv:1808.06601; version 2). arXiv. http://arxiv.org/abs/1808.06601

Wasi, A. T., Šerbetar, K., Islam, R., Rafi, T. H., & Chae, D.-K. (2023). *ARBEx: Attentive feature extraction with reliability balancing for robust facial expression learning* (arXiv:2305.01486). arXiv. https://doi.org/10.48550/arXiv.2305.01486

Wen, Z., Lin, W., Wang, T., & Xu, G. (2023). Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics*, *8*(2), article 2. https://doi.org/10.3390/biomimetics8020199

Whitesell, N. R., & Harter, S. (1996). The interpersonal context of emotion: Anger with close friends and classmates. *Child Development, 67*(4), 1345–1359. https://doi.org/10.1111/j.1467-8624.1996.tb01800.x

Wood, A., Rychlowska, M., Korb, S., & Niedenthal, P. (2016). Fashioning the face: Sensorimotor simulation contributes to facial expression recognition. *Trends in Cognitive Sciences, 20*(3), 227–240. https://doi.org/10.1016/j.tics.2015.12.010

Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision – ECCV 2018 (Vol. 11210, pp. 472–487)*. Springer International Publishing. https://doi.org/10.1007/978-3-030-01231-1_29

Xu, Y., Zhang, J., Zhang, Q., & Tao, D. (2022). ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation (arXiv:2204.12484). arXiv. https://doi.org/10.48550/arXiv.2204.12484

Xu, Z., Hong, Z., Ding, C., Zhu, Z., Han, J., Liu, J., & Ding, E. (2022). *MobileFaceSwap: A lightweight framework for video face swapping* (arXiv:2201.03808; version 1). arXiv. http://arxiv.org/abs/2201.03808

Xue, F., Wang, Q., Tan, Z., Ma, Z., & Guo, G. (2022). Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*, 1–13. https://doi.org/10.1109/TAFFC.2022.3226473

Yang, G. R., & Molano-Mazón, M. (2021). Towards the next generation of recurrent network models for cognitive neuroscience. *Current Opinion in Neurobiology, 70*, 182–192. https://doi.org/10.1016/j.conb.2021.10.015

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences, 45*, e1. https://doi.org/10.1017/S0140525X20001685

Yu, C., Lu, H., Hu, N., Yu, M., Weng, C., Xu, K., Liu, P., Tuo, D., Kang, S., Lei, G., Su, D., & Yu, D. (2019). *DurIAN: Duration informed attention network for multimodal synthesis* (arXiv:1909.01700; version 2). arXiv. http://arxiv.org/abs/1909.01700

Zane, E., Yang, Z., Pozzan, L., Guha, T., Narayanan, S., & Grossman, R. B. (2019). Motion-capture patterns of voluntarily mimicked dynamic facial expressions in children and adolescents with and without ASD. *Journal of Autism and Developmental Disorders, 49*(3), 1062–1079. https://doi.org/10.1007/s10803-018-3811-7