Research

Big team science reveals promises and limitations of machine learning efforts to model physiological markers of affective experience

Nicholas A. Coles[1], Bartosz Perz[2], Maciej Behnke[3], Johannes C. Eichstaedt[4], Soo Hyung Kim[5], Tu N. Vu[5], Chirag Raman[6], Julian Tejada[7], Van-Thong Huynh[8], Guangyi Zhang[9], Tanming Cui[10], Sharanyak Podder[11], Rushi Chavda[12], Shubham Pandey[12], Arpit Upadhyay[12], Jorge I. Padilla-Buritica[13], Carlos J. Barrera Causil[13], Linying Ji[14], Felix Dollack[15], Kiyoshi Kiyokawa[15], Huakun Liu[15], Monica Perusquia-Hernandez[15], Hideaki Uchiyama[15], Xin Wei[15], Houwei Cao[16], Ziqing Yang[16], Alessia Iancarelli[17], Kieran McVeigh[17], Yiyu Wang[17], Isabel M. Berwian[18], Jamie C. Chiu[18], Dan-Mircea Mirea[18], Erik C. Nook[18], Henna I. Vartiainen[18], Claire Whiting[18], Young Won Cho[19], Sy-Miin Chow[19], Zachary F. Fisher[19], Yanling Li[19], Xiaoyue Xiong[19], Yuqi Shen[19], Enzo Tagliazucchi[20], Leandro A. Bugnon[21], Raydonal Ospina[22], Nicolas M. Bruno[23], Tomas A. D'Amelio[23], Federico Zamberlan[23,24], Luis R. Mercado Diaz[25], Javier O. Pinzon-Arenas[25], Hugo F. Posada-Quintero[25], Maneesh Bilalpur[26], Saurabh Hinduja[27], Fernando Marmolejo-

Ramos[28], Shaun Canavan[29], Liza Jivnani[29] and Stanisław Saganowski[2]

[1]University of Florida, Gainesville, FL, USA
[2]Wrocław University of Science and Technology, Wrocław, Województwo Dolnośląskie, Poland
[3]Adam Mickiewicz University, Poznan, Poland
[4]Stanford University, Stanford, CA, USA
[5]Chonnam National University, Gwangju, Jeollanam-do, Republic of Korea
[6]Delft University of Technology, Delft, Zuid-Holland, The Netherlands
[7]Federal University of Sergipe, Sao Cristovao, Sergipe, Brazil
[8]FPT University, Hanoi, Vietnam
[9]Harvard Medical School, Boston, MA, USA
[10]Independent Researcher, State College, PA, USA
[11]Indian Institute of Science Education and Research Bhopal, Bhopal, Madhya Pradesh, India
[12]Indian Institute of Technology Bombay, Mumbai, Maharashtra, India
[13]Institución Universitaria ITM, Medellín, Colombia
[14]Montana State University, Bozeman, MT, USA
[15]Nara Institute of Science and Technology, Ikoma, Nara, Japan
[16]New York Institute of Technology, Old Westbury, NY, USA
[17]Northeastern University—Boston Campus, Boston, MA, USA
[18]Princeton University, Princeton, NJ, USA
[19]The Pennsylvania State University, University Park, PA, USA
[20]Universidad Adolfo Ibanez, Penalolen, Chile
[21]Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET, Santa Fe, Argentina
[22]Universidade Federal da Bahia, Salvador, Brazil
[23]University of Buenos Aires, Buenos Aires, Argentina
[24]Tilburg University, Tilburg, Netherlands
[25]University of Connecticut, Storrs, CT, USA
[26]University of Pittsburgh, Pittsburgh, PA, USA
[27]University of Akron, Akron, OH, USA
[28]Flinders University, Adelaide, South Australia, Australia
[29]University of South Florida, Tampa, FL, USA

NAC, 0000-0001-8583-5610; MB, 0000-0002-2455-4556; V-TH, 0000-0002-6240-2959; HU, 0000-0002-6119-1184; HIV, 0000-0002-5416-898X; ET, 0000-0003-0421-9993; RO, 0000-0002-9884-9090; HFP-Q, 0000-0003-4514-4772

Researchers are increasingly using machine learning to study physiological markers of emotion. We evaluated the promises and limitations of this approach via a big team science competition. Twelve teams competed to predict self-reported affective experiences using a multi-modal set of peripheral nervous system measures. Models were trained and tested in multiple ways: with data divided by participants, targeted emotion, inductions, and time. In 100% of tests, teams outperformed baseline models that made random predictions. In 46% of tests, teams also outperformed baseline models that relied on the simple average of ratings from training datasets. More notably, results uncovered a methodological challenge: multiplicative constraints on generalizability. Inferences about the accuracy and theoretical implications of machine learning efforts depended not only on their architecture, but also how they were trained, tested, and evaluated. For example, some teams performed better when tested on observations from the same (vs. different) subjects seen during training. Such results could be interpreted as evidence against claims of universality. However, such conclusions would be premature because other teams exhibited the opposite pattern. Taken together, results illustrate how big team science can be leveraged to understand the promises and limitations of machine learning methods in affective science and beyond.

1. Introduction

All throughout their lives, people experience phenomenological states they call 'emotions' [1]. For at least a century, researchers have been interested in the extent to which these emotional states have physiological markers. For an emotion artificial intelligence (AI) industry recently valued at $20+ billion [2], these physiological markers may be the key to creating technology that can unobtrusively track and respond to humans' emotions. For emotion theorists, these physiological markers may

provide fundamental insights into the nature of emotions themselves. For example, many theorists posit that emotional experience has a physiological basis, often suggesting that it is partially or fully built off afferent feedback from the peripheral nervous system [3–18]. If true, such accounts provide clues not only about *how* emotional experiences arise but also about *why* such a capacity evolved (e.g. to monitor and regulate physiological states) [4,19].

Over the past couple of decades, researchers have increasingly turned to machine learning methods to capture, study and debate the nature of emotion physiology [20–25]. In the present work, we evaluated the promises and limitations of this approach via a big team science competition. The competition focused on modelling physiological markers of the most elementary component of emotional experience: core affective feelings of valence (pleasantness versus unpleasantness) and arousal (often defined as strength of physiological activity or level of energy) [26]. By crowdsourcing this task, we sought to efficiently identify and probe particularly promising approaches to capturing links between physiology and affective experiences. We further sought to do so in a manner that establishes a high degree of *commensurability* (i.e. comparability). To do so, we had researchers commit to the same sets of training, testing and evaluation procedures. Simultaneously, we examined the impact of non-commensurability by systematically varying many of these methodological decisions— i.e. having researchers engage with multiple training, testing and evaluation frameworks.

Our examination of commensurability was motivated by reviews indicating a general *lack* of commensurability in many machine learning efforts focused on emotion physiology [20]. As an illustrative example, consider the challenges of comparing two simple but influential studies by Picard *et al*. [27] and Haag *et al*. [28] (table 1).

On one hand, Picard *et al*. [27] (i) used peripheral nervous system data from a single participant who completed an emotion self-elicitation task (e.g. guided imagery) to (ii) predict which of eight emotions were targeted by the task. They (iii) defined accuracy as the percentage of times the model correctly predicted which emotion was targeted, and (iv) tested their model using leave-one-out cross-validation. In this validation approach, data are divided into multiple non-overlapping sets that are iteratively used to train and test models. Using this approach, Picard *et al*. [27] concluded that their model achieved 81% classification accuracy.

On the other hand, Haag *et al*. [28] (i) used peripheral nervous system data from a single participant who viewed emotional photos, to (ii) predict how other participants rated those photos. They (iii) defined accuracy as the percentage of times the predicted value fell within a range of ratings from other participants, and (iv) tested their model using hold-out validation. In this validation approach, data are divided into two sets, one for training and one for testing. Using this approach, Haag *et al*. [28] concluded 90% and 97% accuracy for predicted valence and arousal reports, respectively.

In this illustrative example, it is tempting to conclude that Haag *et al*. improved upon Picard *et al*.'s work—or, more broadly, that the field's methods for capturing links between emotion and physiology are improving. However, such claims are difficult to substantiate owing to a lack of commensurability: the two teams used different datasets, outcomes, performance benchmarks and model validation procedures [20,29]. It is certainly possible that Haag *et al*. [28] developed a superior model. However, it is also possible that they studied a simpler emotional context (emotional photos versus self-elicitation), focused on a simpler outcome (core affect prediction versus discrete emotion classification), used a more liberal benchmark (falling within a range of observer ratings versus identifying the targeted emotion) and/or used more rigorous testing procedures (hold out versus leave-one-out cross-validation).

If incommensurability makes it difficult to compare the performance of two machine learning efforts, it is perhaps not surprising that it also confounds attempts to connect such work to theory. Researchers are increasingly using machine learning methodology to weigh in on century-old debates about emotion, e.g. whether underlying patterns are biologically innate, emotion specific, similar across contexts and stable throughout the course of an emotional event [30]. Yet, one possibility is that machine learning researchers reach different conclusions about connections to theory simply because they approach the questions in different ways [31]. The present work evaluates such a possibility by systematically varying training, testing and evaluation procedures in a machine learning challenge involving 12 groups of researchers.

Table 1. Past efforts to use machine learning to predict emotion reports from physiology are difficult to compare. (For example, Picard *et al.* [27] and Haag *et al.* [28] used different datasets, outcomes, benchmarks and testing procedures. The present work standardizes and/or systematically varies those methodological details.)

citation	dataset	outcome	benchmark	testing procedure
Picard *et al*. [27]	*n* = 1 subject completes emotion self-elicitation task	emotion targeted by the self-elicitation task	percentage of times the model correctly predicted which emotion was targeted	leave-one-out cross-validation
Haag *et al*. [28]	*n* = 1 subject views emotional photos	external ratings of emotional photos	percentage of times predicted value fell within a range of the external ratings	hold out validation
present work with 12 teams	*n* = 30 subjects view emotional videos	self-reported affect	absolute prediction error (compared to two baseline models)	across-subject, -emotion, -induction and -time validation

2. Methods

For the competition, we used the Continuously Annotated Signals of Emotion dataset [32]. This dataset contains moment-to-moment measures of affective experience and multi-modal physiology from 30 volunteer subjects recruited from the Institute of Robotics and Mechatronics in Germany (15 males, age 28.6 ± 4.8 years and 15 females, age 25.7 ± 3.1 years). These subjects encountered two video inductions of fear, boredom, relaxation and amusement. In total, there were eight inductions, each 2–3 min in length. These videos were initially selected based on their use in prior research, but participants' affect reports provided further evidence of their effectiveness as emotion inductions [32].

Throughout the video inductions, participants used a joystick to navigate a continuous, 9-point, two-dimensional grid measuring valence (i.e. pleasantness versus unpleasantness) on the *x*-axis and arousal (i.e. low energy versus high energy) on the *y*-axis. Simultaneously, multi-modal peripheral nervous system measurements were collected: electrocardiography, blood volume pulse, electrodermal activity, respiration, skin temperature and electromyography activity over the zygomaticus major, corrugator supercilii and trapezius muscles. Notably, both the affect reports and peripheral nervous system signals were collected moment-to-moment (20 and 1000 samples s^{-1}, respectively). This provided teams with a large set of observations with high temporal resolution. For each participant, the original dataset contained approximately (i) 25 401 valence and arousal reports and (ii) 1 270 083 recordings of peripheral nervous system activity. Using these observations, teams were challenged to use peripheral nervous system features to predict affect reports via their preferred machine learning methodology.

2.1. Teams

For the competition, we sought to recruit up to 15 teams based on the availability of funds. Teams were eligible to participate if they agreed to (i) make their code openly available, (ii) collaborate on a manuscript describing the challenge results, and (iii) not cheat (e.g. look for the original dataset). Eighteen teams completed the application to join the challenge, which asked them to report previous experience with machine learning challenges, the number of papers published in this domain, their planned approach to the challenge and the CV of members of their teams. The competition organizers selected teams based on averaged ratings of their (i) experience with machine learning challenges (N.A.C.), (ii) expertise in emotion research (M.B.) and/or (iii) planned approach to the challenge (S.S.). These evaluations were made independently and then discussed as a group. Of the 15 teams that were invited to compete, three dropped out owing to difficulties encountered during the challenge. In total, 12 teams completed the challenge, using a variety of modelling approaches (electronic supplementary material, table S1) [33–35].
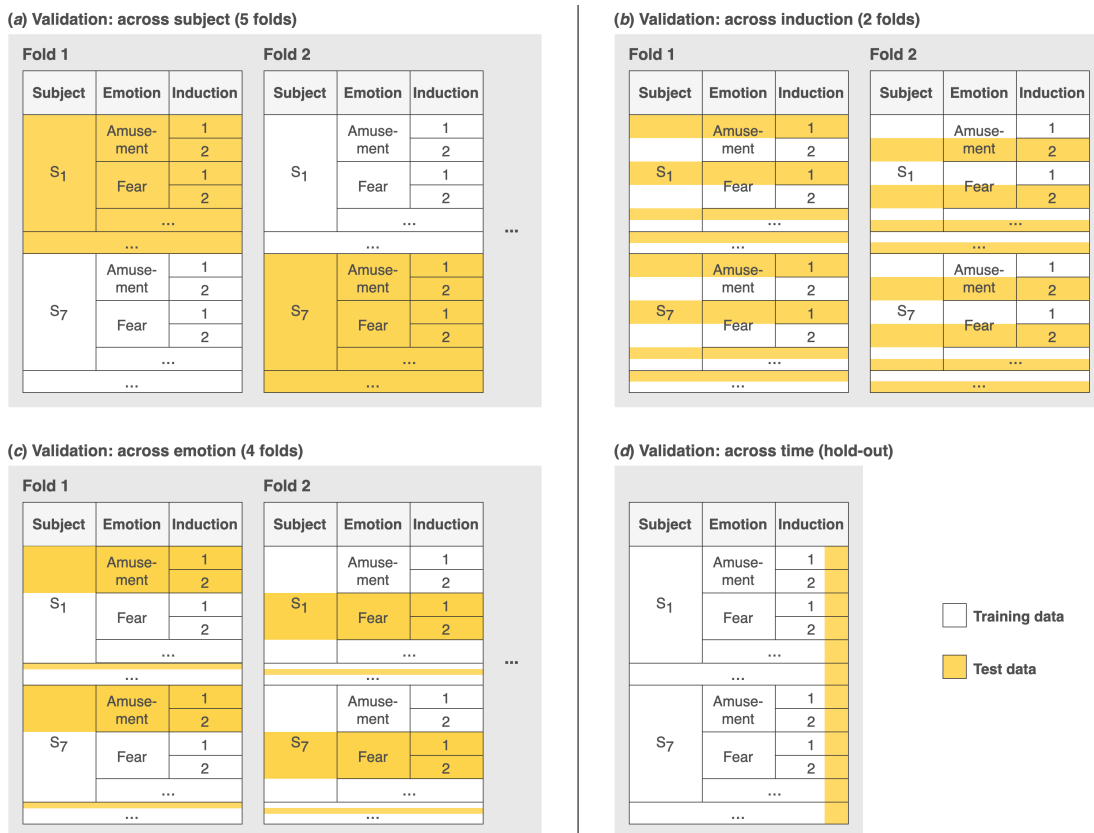
Figure 1. Overview of the four validation approaches. (*a-c*) Data were divided into subsets (folds). Models were iteratively trained on one set of folds (white) and tested on the remaining fold (yellow). (*d*) Models were trained on the beginning of all inductions and tested on later parts of the inductions.

Teams received $300 for completing the competition. Those who developed approaches judged to be the most promising received a $200 bonus. This payment structure was designed to balance incentives for effort ($300 for all teams) and performance ($200 for most promising approaches). To subjectively identify the most promising approaches, two competition organizers (S.S. and B.P.) worked together to examine (i) how well the teams performed by inspecting their root mean square error, and (ii) if top performing, their underlying code. This two-step process allowed us to consider performance, perceived methodological soundness and reproducibility when choosing which models to further evaluate. This allowed us to efficiently select promising models for follow-up investigations: a selection process that helps conserve resources consumed by large-scale machine learning efforts [36].

2.2. Model training and testing

There were no constraints on how teams were instructed to model the data. For example, some teams deployed relatively simply tree-based ensemble models, whereas others deployed relatively complex deep learning techniques (electronic supplementary material, table S1).

To explore the implications of decisions about model training and testing, we examined four different approaches (figure 1):

(i) for *across-subject validation*, we used leave-*N*-subjects-out validation (figure 1*a*). Participants were randomly divided into five folds. Teams trained models on four folds and tested models on a fifth fold. This was repeated for each combination of folds;

(ii) for *across-emotion validation*, we used a leave-one-emotion-out validation approach (figure 1*c*). As a reminder, four emotions were targeted (via videos) in the original dataset: amusement, fear, boredom and relaxation. We created one fold for each targeted emotion (four folds total). Teams trained models on data from three targeted emotions and tested models on data from a fourth targeted emotion. This was repeated for each combination of folds;

(iii) for *across-induction validation*, we used a leave-one-video-out validation approach (figure 1*b*). As a reminder, each targeted emotion was induced through two different videos. For each targeted emotion, teams trained models on data from one video and tested models on data from the second video; and

(iv) for *across-time validation*, we used a hold-out validation approach focused on chronology (figure 1*d*). For each participant, we divided the data from each emotion induction into training and test sets based on time. Teams trained models on data from the beginning of the inductions and tested the models on data from the later parts of the inductions.

Notably, the four validation approaches represent what researchers might do to evaluate theoretical debates about the extent to which links between peripheral nervous system activity and affective experience are biologically innate (across-subject validation), emotion specific (across-emotion validation), similar across contexts (across-induction validation) and stable throughout the course of an emotional event (across-time validation) [30].

Test data files contained 30 s of affect reports (removed during testing) and their corresponding physiological recordings. Test data files also contained 10 s of physiological recordings that preceded and succeeded the affect report window. This allowed teams to potentially use short periods of past and future physiological data to predict affect reports. Teams were permitted to build different models for different validation approaches. Before submitting their models for evaluation on the final set of test data, teams were allowed to conduct up to three preliminary tests on a subset (50%) of test data. This approach is often used in machine learning to prevent overfitting on training data, thus promoting the generalizability of the models. Eight teams conducted these preliminary tests (see the electronic supplementary material, table S1).

The outcome of interest was the absolute value of the prediction error for self-reports of valence and arousal. We chose to focus on self-reports because it most closely maps onto a central construct of interest in emotion research: the *experience* of affective states. Prediction error was estimated separately for valence and arousal reports and then summed. To examine the impact of decisions about benchmarking, we compared teams to two simple baseline models: (i) a *random baseline model* that made random (within the range of the measure) predictions about self-reported valence and arousal reports; and (ii) a *mean baseline model* that uniformly predicted each rating based on the observed mean in the training dataset. However, other baseline models could certainly be considered, such as ones that calculate tailored subject-specific mean ratings, preserve first-order signal statistics and/or constrain continuity, temporal variability and autocorrelation.

For the three models judged to be the most promising, follow-up tests were conducted to evaluate the extent to which the models used peripheral nervous system features. For these tests, the peripheral nervous system data in the testing files were replaced with simulated physiological randomness: $N(\mu = 0, \sigma = 1)$.

3. Results

To examine each team's accuracy, we used mixed-effect regression. Mixed-effect regression is a robust analytic approach that can be used to accommodate non-independent observations, which often lead to inflated error rates in traditional linear regression or data loss when averaged to avoid non-independence [37–39]. For each team, we regressed absolute prediction error as a function of: (i) whether the prediction came from a model developed by the team, the random baseline model or the mean baseline model; (ii) the validation approach used for testing; (iii) a higher-order interaction between model source and validation approach; and (iv) random intercepts for each subject and video in the dataset. Random intercepts were included to accommodate non-independent observations from the same participants and videos. For each validation approach, we used model-derived pairwise contrasts to estimate and test the significance of the mean difference (MD) in the absolute prediction error between the model developed by the team and each baseline model.

As illustrated in figure 2*a*, inferences about how well teams modelled affect reports depend on which baseline model is considered. When compared to a *random baseline model*, every team in every validation scenario had more accurate predictions ($1.90 < MD > 0.48$, all $z > 130.76$, all $p < 0.001$). However, teams did not always make more accurate predictions than a *mean baseline model*. Seven teams (58%) outperformed the mean baseline model in across-subject ($0.50 < MD > 0.008$, all $z > 2.54$, all $p < 0.05$) and across-time validation ($0.74 < MD > 0.018$, all $z > 5.36$, all $p < 0.001$); five teams (42%)

outperformed the mean baseline model in across-emotion validation (0.71 < MD > 0.02, all z > 6.01, all p < 0.001); and three (25%) teams outperformed the mean baseline model in across-induction validation (0.52 < MD > 0.23, all z > 64.20, all p < 0.001).

As further illustrated in figure 2*b*, the manner in which prediction accuracy varied across validation approaches differed across teams. For example, team 1 achieved lower prediction error in across-subject versus across-time validation. Team 4 exhibited the opposite pattern. These results highlight how constraints on generalizability are *multiplicative*. Inferences depended on the interactive effect of decisions about (i) what models to deploy and (ii) how to evaluate those models.

3.1. Further evidence of the role of peripheral nervous system activity

The above results are consistent with previous claims that machine learning models can capture links between peripheral nervous system activity and affect reports [40,41]. However, although the models always outperformed random guessing, they did not always outperform a mean baseline model that uniformly predicted affect ratings as the mean of the ratings observed in the training dataset (figure 1*a*). This raises questions about the extent to which the models' predictions were driven by the recovery of (i) theory-relevant links between peripheral nervous system activity and affect reports versus (ii) theory-irrelevant averages of affect reports in the training dataset.[1]

One benefit of big team science is that it can be used to efficiently identify and probe particularly promising models. To further investigate the role of peripheral nervous system activity, competition organizers reviewed the openly available code for top-performing submissions and [subjectively] chose three that seemed particularly promising. The teams' models were then re-tested on the same data with one change: measures of peripheral nervous system activity were replaced with simulated physiological randomness (figure 3*a*).

Using mixed-effect regression, we regressed each of the three team's absolute prediction error as a function of (i) the validation approach used for testing, (ii) whether the test data contained real or simulated physiological randomness, (iii) a higher-order interaction between validation approach and whether the physiology input was real, and (iv) random intercepts for each subject and video in the dataset. In all but one case (MD = 0.00, z = −1.07, p = 0.29), the teams' prediction accuracy decreased when tested on simulated physiological randomness (−0.33< MD > −0.018, all z < −6.96, all p < 0.001; figure 3). In other words, accuracy decreased in 93% of tests with simulated physiological randomness. These results provide evidence that most—but not all—high-performing models actually relied on the provided peripheral nervous system signals to predict affect reports. However, results also highlight the limits of what models learned from physiology, with real physiology only improving affect report predictions by a maximum of 0.33 points (on a 9-point scale).

4. Discussion

Taken together, the results of our big team effort reveal both the promises and limitations of machine learning efforts to model potentially complex physiological markers of affective experience. In four tests, models developed by 12 teams of researchers uniformly achieved higher accuracy than what would be expected by mere random guessing. About half the time, accuracy was also higher than what would be expected by merely using mean ratings from training datasets. Further evidence of the role of peripheral nervous system activity comes from follow-up tests with simulated physiological randomness, which nearly uniformly caused models from three selected teams to become less accurate. However, the magnitude of differences in prediction accuracy in many tests was small, illustrating extensive opportunity for future improvements.

---

[1]One challenge with connecting machine learning to theoretical debates about emotion physiology is model interpretability; i.e. it is unclear if successful predictions are driven by (i) theory-relevant links between peripheral nervous system and affect reports, and (ii) theory-irrelevant averages of affect reports in the training dataset.To illustrate this point, consider a classic linear regression context. Even without modelling physiology-related parameters, an intercept-only regression can often make reasonable predictions about participants' affect ratings (if mean affect ratings captured by the model intercept are similar in training and test data). To more specifically make claims about physiology, linear links between physiology parameters and affect ratings would be formally modelled and inspected.Unlike linear regression, many machine learning methods (e.g. deep learning) do not fit interpretable parameters. Thus, researchers often need to deploy supplemental investigations to substantiate claims about what their models learned (if anything) about emotion physiology.
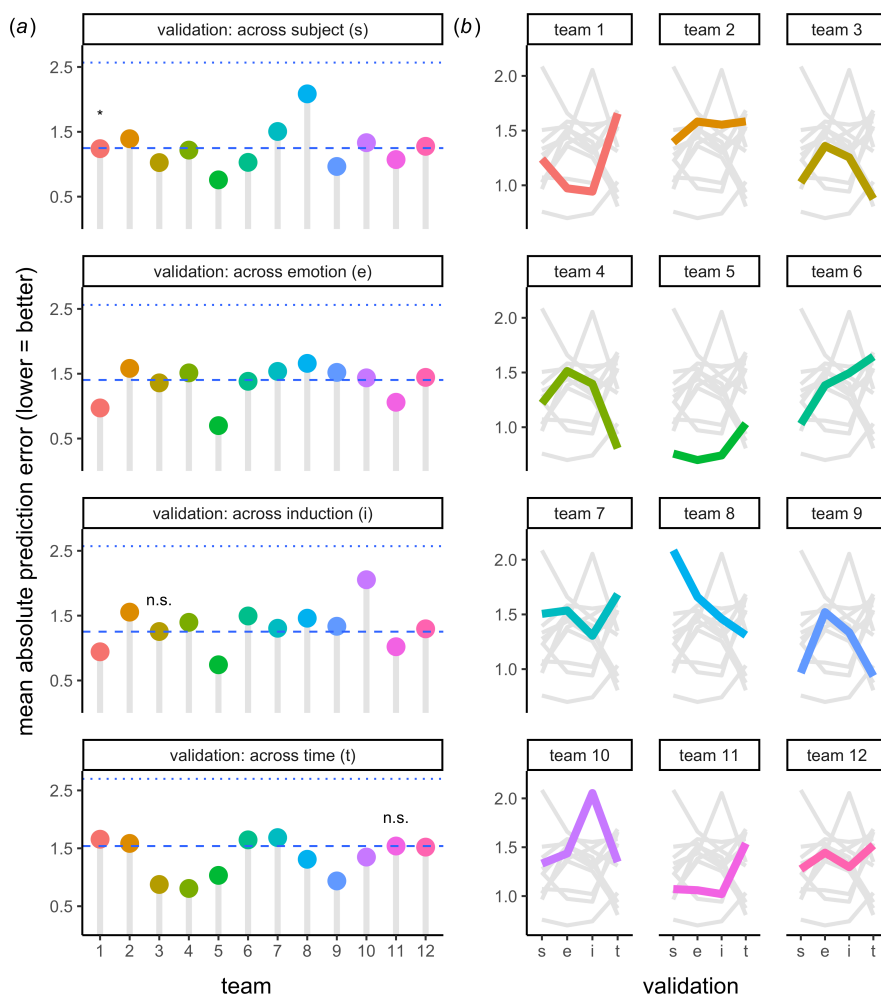
Figure 2. Differences in how machine learning models are developed, benchmarked and tested interactively shape conclusions about their ability to predict affect reports from physiology. (*a*) Absolute error of affect report predictions (*y*-axis) made by 12 teams' models (*x*-axis). Validation approaches (panels) are visualized separately. Models are compared to a random baseline (upper dotted line) and mean baseline (lower dashed line). Note: $p < 0.001$ unless otherwise indicated; an asterisk denotes $p < 0.05$ for mean baseline comparison; *n.s.* denotes $p > 0.05$ for mean baseline comparison. (*b*) Re-illustration of teams' (panel) prediction error (*y*-axis) across validation approaches (*x*-axis).

More centrally, however, our results underscore challenges that past and future researchers face with commensurability and generalizability in research that seeks to use machine learning to predict and understand emotion [42,43]. Our results indicate that differences in how models are developed, benchmarked *and* tested can impact researchers' conclusions. For instance, focusing on a random baseline (versus mean baseline) leads to a more optimistic interpretation of models' accuracy—as does focusing on across-time validation (versus, e.g. across-emotion validation). These results have implications for ongoing discussions about the potential benefits (e.g. unobtrusive measurement of internal emotional states) and harms (e.g. inaccurate predictions) of emotion recognition technologies [44–47], such as those being pursued by a $20+ billion emotion AI industry [2].

Even more challenging is our observation that constraints on generalizability can be *multiplicative* (i.e. interactive) [43,48]. For example, the accuracy of predictions depended both on the modelling and validation approaches. However, it is inadvisable to make broad conclusions about the impact of any one of these decisions because they appeared to have an *interactive* effect on prediction accuracy. For example, team 1 achieved lower prediction error in across-subject versus across-time validation. Some researchers may be tempted to conclude that this provides evidence that links between peripheral nervous system activity and affect reports vary more within- versus between-persons. This conclusion, if true, would bolster claims that such links are biologically innate [49], but perhaps sensitive to context. This conclusion would also suggest that future research should focus less on the diversity of the sampled population and more on the diversity of the emotional contexts they encounter. However,
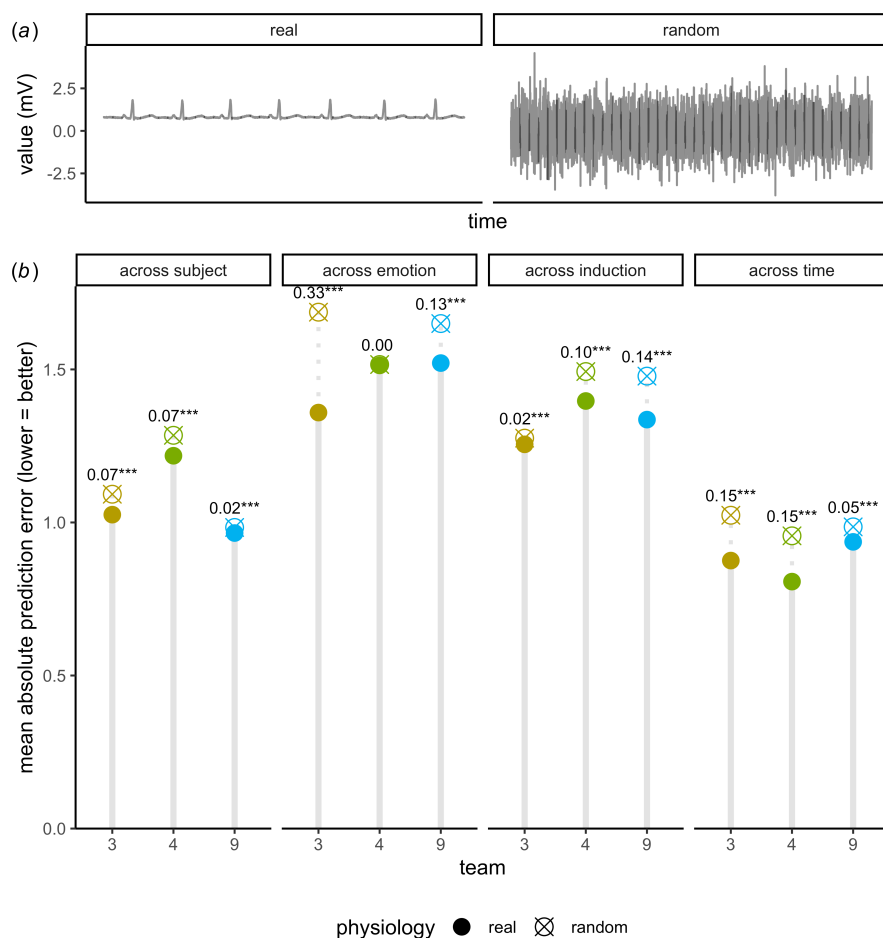
Figure 3. Three re-examined models typically exhibited less accuracy when re-tested on simulated physiological randomness. (*a*) Example of real electrocardiography signal versus simulated physiological randomness. (*b*) Absolute error of affect report predictions (*x*-axis) made by models developed by three teams (*y*-axis) when tested on real physiology (circle) versus simulated physiological randomness (crossed circle). Results are visualized separately for four validation approaches (panels). Significance levels correspond to mean differences in performance when a model was tested on real versus simulated physiological randomness. Note: ***$p < 0.001$.

such conclusions are premature when considering that differences in both modelling *and* validation approaches have *interactive* effects on prediction accuracy. Indeed, models developed by other teams (e.g. team 8) exhibited the opposite pattern: that prediction error is lower in across-time versus across-subject validation. Such results could be interpreted as evidence *against* claims of biological innateness and would underscore the importance of collecting diverse participant samples [50].

Multiplicative constraints on generalizability will be important to keep in mind as researchers increasingly use machine learning not only to predict emotion—but also to evaluate theoretical claims about its nature [31,51,52]. Our work focused on a specific theoretical issue in a specific methodological context: whether there are detectable links between physiological states and core affective experiences among participants in Germany in a stationary and controlled laboratory context. However, we suspect that our observed multiplicative constraints on generalizability will apply to other theoretical debates and methodological contexts in affective science. This includes debates about the number and discreteness of emotion categories [53,54], the extent to which their physiological correlates are invariant across different contexts and people [51], which physiological features (if any) are most robustly associated with specific emotional states [54,55], the utility of using wearable sensors to track emotional processes in non-laboratory contexts [20,56], and the performance of emotion recognition models when exposed to entirely new contexts (e.g. contexts with completely new data,[2] such as real world contexts with ambulatory measures of physiology).

---

[2]When the competition was designed, the organizers searched for an external dataset with similar methodological characteristics (moment-to-moment measures of core affect and a similar suite of physiological measures). Such efforts were not successful (even after reaching out to the developers of the dataset that was used).

Although our work highlights challenges with commensurability and multiplicative constraints on generalizability, it also provides proof-of-concept for a potential methodological response: big team science [57–61]. Big team science effectively allowed us to use the wisdom-of-crowds to evaluate a fundamental theoretical question in affective science. Standardizing methodological decisions about data sources, benchmarks and testing procedures permitted cleaner comparisons of teams' different approaches. Further introducing *systematic variation* in specific methodological decisions (e.g. testing procedures) allowed us to empirically examine the extent to which these decisions constrain the generalizability of inferences. Finally, requiring that teams make their materials and code openly available allowed us (and future researchers) to further inspect teams' models, reproduce their solutions and identify approaches that seem most promising for follow-up research [62]. For example, although our analyses did not investigate the impact of specific methodological decisions, this remains a promising goal for future research.

Our examination of the use of big team science in machine learning research on emotion also yielded lessons about ways that future collaborative efforts can be improved and expanded. For instance, feasibility constraints prohibited the competition organizers from closely examining the code for all submissions, performing comprehensive evaluations of model behaviour and working closely with teams to navigate potential disagreements about the appropriateness of their methodology [63]. Expanding upon current guidelines for crowdsourced model development [64], developing protocols and best practices for *peer code review* could have better enabled the crowdsourcing of this task [65–68]. Such protocols, for example, could help researchers examine whether specific decisions predict variability in the results of machine learning efforts [69]. Teams that entered the competition also faced resource constraints, with many expressing that they would have benefited from (i) more time to work on the challenge, (ii) datasets with more peripheral nervous system measures, observations and cultural variability, and/or (iii) access to more powerful computing resources. Recent pushes to collaborate on dataset development [57–60] and provide access to shared computing resources [70] may prove instrumental in helping researchers overcome those barriers.

Despite existing constraints, our work raises an exciting possibility: advancements in machine learning and collaborative research methods provide researchers with new tools for tackling ultra-complex questions in affective science and beyond. However, when doing so, researchers will have to grapple with multiplicative constraints on generalizability.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. Data, materials and code are openly available at [71].

Supplementary material is available online [72].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. N.A.C.: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, visualization, writing—original draft, writing—review and editing; B.P.: conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualization, writing—original draft, writing—review and editing; M.B.: conceptualization, data curation, investigation, methodology, project administration, validation, visualization, writing—original draft, writing—review and editing; J.C.E.: conceptualization, funding acquisition, visualization, writing—original draft, writing—review and editing; S.H.K.: investigation, methodology, writing—review and editing; T.N.V.: investigation, methodology, writing—review and editing; C.R.: investigation, methodology, writing—review and editing; J.T.: investigation, methodology, writing—review and editing; V.-T.H.: investigation, methodology, writing—review and editing; G.Z.: investigation, methodology, writing—review and editing; T.C.: investigation, methodology, writing—review and editing; S.P.: investigation, methodology, writing—review and editing; R.C.: investigation, methodology, writing—review and editing; S.P.: investigation, methodology, writing—review and editing; A.U.: investigation, methodology, writing—review and editing; J.I.P.-B.: investigation, methodology, writing—review and editing; C.J.B.C.: investigation, methodology, writing—review and editing; L.J.: investigation, methodology, writing—review and editing; F.D.: investigation, methodology, writing—review and editing; K.K.: investigation, methodology, writing—review and editing; H.L.: investigation, methodology, writing—review and editing; M.P.-H.: investigation, methodology, writing—review and editing; H.U.: investigation, methodology, writing—review and editing; X.W.: investigation, methodology, writing—review and editing; H.C.: investigation, methodology, writing—review and editing; Z.Y.: investigation, methodology, writing—review and editing; A.I.: investigation, methodology, writing—review and editing; K.McV.: investigation, methodology, writing—review and editing; Y.W.: investigation, methodology, writing—review and editing; I.M.B.: investigation, methodology, writing—review and editing; J.C.C.: investigation, methodology, writing—review and editing; D.-M.M.: investigation, methodology, writing—review and editing; E.C.N.: investigation, methodology, writing—review and editing; H.I.V.: investigation, methodology, writing—review and editing; C.W.: investigation, methodology, writing—review and editing; Y.W.C.: investigation, methodology, writing—review and editing; S.-M.C.: investigation, methodology, writing—review and editing; Z.F.F.: investigation, methodology, writing—review and editing; Y.L.: investigation, methodology, writing—review and editing; X.X.: investigation, methodology, writing—review and editing; Y.S.: investigation, methodology, writing—review and editing; E.T.: investigation, methodology, writing—review and editing;

11

L.A.B.: investigation, methodology, writing—review and editing; R.O.: investigation, methodology, writing—review and editing; N.M.B.: investigation, methodology, writing—review and editing; T.D.: investigation, methodology, writing—review and editing; F.Z.: investigation, methodology, writing—review and editing; L.R.M.D.: investigation, methodology, writing—review and editing; J.O.P.-A.: investigation, methodology, writing—review and editing; H.F.P.-Q.: investigation, methodology, writing—review and editing; M.B.: investigation, methodology, writing—review and editing; S.H.: investigation, methodology, writing—review and editing; F.M.-R.: investigation, methodology, writing—review and editing; S.C.: investigation, methodology, writing—review and editing; L.J.: investigation, methodology, writing—review and editing; S.S.: conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, supervision, validation, visualization, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

References

1. Damasio AR. 1994 Descartes' error and the future of human life. *Sci. Am.* **271**, 144–144. (doi:10.1038/scientificamerican1094-144)

2. Telford T. 2019 'Emotion detection' AI is a $20 billion industry. New research says it can't do what it claims. *Washington Post*. See https://www.washingtonpost.com/business/2019/07/31/emotion-detection-ai-is-billion-industry-new-research-says-it-cant-do-what-it-claims/.

3. Cacioppo JT, Berntson GG, Klein DJ. 1992 What is an emotion? The role of somatovisceral afference, with special emphasis on somatovisceral 'illusions'. In *Review of personality and social psychology: emotion and social behavior* (ed. MS Clark), pp. 63–98. Thousand Oaks, CA: Sage Publications.

4. Damasio A, Carvalho GB. 2013 The nature of feelings: evolutionary and neurobiological origins. *Nat. Rev. Neurosci.* **14**, 143–152. (doi:10.1038/nrn3403)

5. Scherer KR, Moors A. 2019 The emotion process: event appraisal and component differentiation. *Annu. Rev. Psychol.* **70**, 719–745. (doi:10.1146/annurev-psych-122216-011854)

6. Feldman MJ, Bliss-Moreau E, Lindquist KA. 2024 The neurobiology of interoception and affect. *Trends Cogn. Sci.* **28**, 1–19. (doi:10.1016/j.tics.2024.01.009)

7. Barrett LF. 2017 The theory of constructed emotion: an active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* **12**, 1–23. (doi:10.1093/scan/nsw154)

8. Coles NA, Gaertner L, Frohlich B, Larsen JT, Basnight-Brown DM. 2022 Fact or artifact? Demand characteristics and participants' beliefs can moderate, but do not fully account for, the effects of facial feedback on emotional experience. *J. Pers. Soc. Psychol.* **124**, 287–310. (doi:10.1037/pspa0000316)

9. Coles NA *et al*. 2022 A multi-lab test of the facial feedback hypothesis by the many smiles collaboration. *Nat. Hum. Behav.* **6**, 1731–1742. (doi:10.1038/s41562-022-01458-9)

10. Coles NA, Larsen JT, Lench HC. 2019 A meta-analysis of the facial feedback literature: effects of facial feedback on emotional experience are small and variable. *Psychol. Bull.* **145**, 610–651. (doi:10.1037/bul0000194)

11. Tomkins S. 1962 *Affect imagery consciousness, vol. 1: the positive affects*. New York, NY: Springer.

12. Wood A, Rychlowska M, Korb S, Niedenthal P. 2016 Fashioning the face: sensorimotor simulation contributes to facial expression recognition. *Trends Cogn. Sci.* **20**, 227–240. (doi:10.1016/j.tics.2015.12.010)

13. James W. 1884 What is an emotion? *Mind* **9**, 188–205. (doi:10.1037/11304-033)

14. James W. 1894 Discussion: the physical basis of emotion. *Psychol. Rev.* **1**, 516–529. (doi:10.1037/h0065078)

15. Laird JD, Bresler C. 1992 The process of emotional experience: a self-perception theory. In *Review of personality and social psychology: emotion* (ed. MS Clark), pp. 213–234. Newbury Park, CA: Sage Publications.

16. Lange CG. 1885 *Om sindsbevaegelser; et psyko-fysiologisk studie*. Lund, Sweden: Jacob Lunds Forlag.

17. Levenson RW, Ekman P, Friesen WV. 1990 Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology* **27**, 363–384. (doi:10.1111/j.1469-8986.1990.tb02330.x)

18. Russell JA. 1980 A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178. (doi:10.1037/h0077714)

19. Barrett LF. 2017 *How emotions are made: the secret life of the brain.* London, UK: Pan Macmillan.

20. Saganowski S, Perz B, Polak AG, Kazienko P. 2022 Emotion recognition for everyday life using physiological signals from wearables: a systematic literature review. *IEEE Trans. Affect. Comput.* **12**, 1–20. (doi:10.1109/TAFFC.2022.3176135)

21. Barrett LF, Adolphs R, Marsella S, Martinez AM, Pollak SD. 2019 Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychol. Sci. Public Interest* **20**, 1–68. (doi:10.1177/1529100619832930)

22. Wang Y *et al*. 2022 A systematic review on affective computing: emotion models, databases, and recent advances. *Inf. Fusion* **83**, 19–52. (doi:10.1016/j.inffus.2022.03.009)

23. Yadegaridehkordi E, Noor NFBM, Ayub MNB, Affal HB, Hussin NB. 2019 Affective computing in education: a systematic review and future research. *Comput. Educ.* **142**, 103649. (doi:10.1016/j.compedu.2019.103649)

24. Assabumrungrat R, Sangnark S, Charoenpattarawut T, Polpakdee W, Sudhawiyangkul T, Boonchieng E, Wilaiprasitporn T. 2021 Ubiquitous affective computing: a review. *IEEE Sens. J.* **22**, 1867–1881. (doi:10.1109/JSEN.2021.3138269)

25. Cortiñas-Lorenzo K, Lacey G. 2023 Toward explainable affective computing: a review. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 13101–13121. (doi:10.1109/TNNLS.2023.3270027)

26. Russell JA. 2003 Core affect and the psychological construction of emotion. *Psychol. Rev.* **110**, 145–172. (doi:10.1037//0033-295x.110.1.145)

27. Picard RW, Vyzas E, Healey J. 2001 Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**, 1175–1191. (doi:10.1109/34.954607)

28. Haag A, Goronzy S, Schaich P, Williams J. 2004 Emotion recognition using bio-sensors: first steps towards an automatic system. In *Lecture notes in computer science affective dialogue systems* (eds E André, L Dybkjær, W Minker, P Heisterkamp), pp. 36–48. Berlin, Germany: Springer. (doi:10.1007/978-3-540-24842-2_4)

29. Calvo RA, D'Mello S. 2010 Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.* **1**, 18–37. (doi:10.1109/t-affc.2010.1)

30. Lindquist KA, Siegel EH, Quigley KS, Barrett LF. 2013 The hundred-year emotion war: are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011). *Psychol. Bull.* **139**, 255–263. (doi:10.1037/a0029038)

31. Azari B *et al*. 2020 Comparing supervised and unsupervised approaches to emotion categorization in the human brain, body, and subjective experience. *Sci. Rep.* **10**, 1–17. (doi:10.1038/s41598-020-77117-8)

32. Sharma K, Castellini C, van den Broek EL, Albu-Schaeffer A, Schwenker F. 2019 A dataset of continuous affect annotations and physiological signals for emotion analysis. *Sci. Data* **6**, 1–13. (doi:10.1038/s41597-019-0209-0)

33. D'Amelio TA, Bruno NM, Bugnon LA, Zamberlan F, Tagliazucchi E. 2023 Affective computing as a tool for understanding emotion dynamics from physiology: a predictive modeling study of arousal and valence. In *11th Int. Conf. on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, Cambridge, MA, USA, pp. 1–7. New York, NY: IEEE. (doi:10.1109/ACIIW59127.2023.10388155)

34. Pinzon-Arenas JO, Mercado-Diaz L, Tejada J, Marmolejo-Ramos F, Barrera-Causil C, Padilla JI, Ospina R, Posada-Quintero H. 2023 Deep learning analysis of electrophysiological series for continuous emotional state detection. In *2023 11th Int. Conf. on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, *Cambridge, MA, USA*, pp. 1–8. (doi:10.1109/ACIIW59127.2023.10388196)

35. Dollack F, Kiyokawa K, Liu H, Perusquia-Hernandez M, Raman C, Uchiyama H, Wei X. 2023 Ensemble learning to assess dynamics of affective experience ratings and physiological change. In *11th Int. Conf. on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, *Cambridge, MA, USA*, pp. 1–8. New York, NY: IEEE. (doi:10.1109/ACIIW59127.2023.10388116)

36. Dhar P. 2020 The carbon impact of artificial intelligence. *Nat. Mach. Intell.* **2**, 423–425. (doi:10.1038/s42256-020-0219-9)

37. Brauer M, Curtin JJ. 2018 Linear mixed-effects models and the analysis of nonindependent data: a unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychol. Methods* **23**, 389–411. (doi:10.1037/met0000159)

38. Schielzeth H *et al*. 2020 Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol. Evol.* **11**, 1141–1152. (doi:10.1111/2041-210x.13434)

39. Frank MC, Braginsky M, Cachia J, Coles NA, Hardwicke TE, Hawkins RD. 2025 *Experimentology: an open science approach to experimental psychology methods*. Boston, MA: MIT Press. (doi:10.25936/3JP6-5M50)

40. Asutay E, Genevsky A, Barrett LF, Hamilton JP, Slovic P, Västfjäll D. 2019 Affective calculus: the construction of affect through information integration over time. *Emotion* **21**, 159–174. (doi:10.1037/emo0000681)

41. Cowen AS, Keltner D, Schroff F, Jou B, Adam H, Prasad G. 2020 Sixteen facial expressions occur in similar contexts worldwide. *Nature* **589**, 251–257. (doi:10.1038/s41586-020-3037-7)

42. Yarkoni T. 2022 The generalizability crisis. *Behav. Brain Sci.* **45**, e1. (doi:10.1017/S0140525X20001685)

43. Almaatouq A, Griffiths TL, Suchow JW, Whiting ME, Evans J, Watts DJ. 2024 Beyond playing 20 questions with nature: integrative experiment design in the social and behavioral sciences. *Behav. Brain Sci.* **47**, 1–55. (doi:10.1017/S0140525X22002874)

44. Hernandez J, Lovejoy J, McDuff D, Suh J, O'Brien T, Sethumadhavan A, Greene G, Picard R, Czerwinski M. 2021 Guidelines for assessing and minimizing risks of emotion recognition applications. In *9th Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, *Nara, Japan*, pp. 1–8. (doi:10.1109/ACII52823.2021.9597452)

45. Ong DC. 2021 An ethical framework for guiding the development of affectively-aware artificial intelligence. In *9th Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, *Nara, Japan*, pp. 1–8. (doi:10.1109/ACII52823.2021.9597441)

46. Bryant D, Howard A. 2021 Age bias in emotion detection: an analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proc. of the 2021 AAAI/ACM Conf. on AI, Ethics, and Society*, pp. 638–644. New York, NY: ACM. (doi:10.1145/3461702.3462609)

47. Boyd KL, Andalibi N. 2023 Automated emotion recognition in the workplace: how proposed technologies reveal potential futures of work. *Proc. ACM Hum. Comput. Interact.* **95**, 1–95. (doi:10.1145/3593053)

48. McVeigh K, Kleckner IR, Quigley KS, Satpute AB. 2023 Fear-related psychophysiological patterns are situation and individual dependent: a Bayesian model comparison approach. *Emotion* **24**, 506–521. (doi:10.1037/emo0001265)

49. Tracy JL, Randles D. 2011 Four models of basic emotions: a review of Ekman and Cordaro, Izard, Ievenson, and Panksepp and Watt. *Emot. Rev.* **3**, 397–405. (doi:10.1177/1754073911410747)

50. Hussein A, Kadir S, Rizhinashvili D, Kuklianov D, Alisinanoglu F, Ofodile I, Ozcinar C, Anbarjafari G. 2023 Ethical AI in facial expression analysis: racial bias. *Signal. Image Video Process.* **17**, 399–406. (doi:10.1007/s11760-022-02246-8)

51. Hoemann K, Khan Z, Feldman MJ, Nielson C, Devlin M, Dy J, Barrett LF, Wormwood JB, Quigley KS. 2020 Context-aware experience sampling reveals the scale of variation in affective experience. *Sci. Rep.* **10**, 1–16. (doi:10.1038/s41598-020-69180-y)

52. Barrett LF. 2021 Debate about universal facial expressions goes big. *Nature* **589**, 200–201. (doi:10.1038/d41586-020-03509-5)

53. Cowen AS, Keltner D. 2017 Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl Acad. Sci. USA* **114**, E7900–E7909. (doi:10.1073/pnas.1702247114)

54. Behnke M, Kreibig SD, Kaczmarek LD, Assink M, Gross JJ. 2022 Autonomic nervous system activity during positive emotions: a meta-analytic review. *Emot. Rev.* **14**, 132–160. (doi:10.1177/17540739211073084)

55. Allport FH. 1922 A physiological-genetic theory of feeling and emotion. *Psychol. Rev.* **29**, 132–139. (doi:10.1037/h0075652)

56. Saganowski S. 2022 Bringing emotion recognition out of the lab into real life: recent advances in sensors and machine learning. *Electronics* **11**, 496. (doi:10.3390/electronics11030496)

57. Coles NA, Hamlin JK, Sullivan LL, Parker TH, Altschul D. 2022 Build up big-team science. *Nature* **601**, 505–507. (doi:10.1038/d41586-022-00150-2)

58. Forscher PS, Wagenmakers EJ, Coles NA, Silan MA, Dutra N, Basnight-Brown D, IJzerman H. 2023 The benefits, barriers, and risks of big-team science. *Perspect. Psychol. Sci.* **18**, 607–623. (doi:10.1177/17456916221082970)

59. Coles NA, DeBruine LM, Azevedo F, Baumgartner HA, Frank MC. 2023 'Big team' science challenges us to reconsider authorship. *Nat. Hum. Behav.* **7**, 665–667. (doi:10.1038/s41562-023-01572-2)

60. Baumgartner HA *et al*. 2023 How to build up big team science: a practical guide for large-scale collaborations. *R. Soc. Open Sci.* **10**, 230235. (doi:10.1098/rsos.230235)

61. Coles NA. 2024 The prevalence, impact, and disruptiveness of big team science. *PsyArXiv* (doi:10.31234/osf.io/q68yv)

62. Haibe-Kains B *et al*. 2020 Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16. (doi:10.1038/s41586-020-2766-y)

63. Coles NA, Tenney ER, Chin JM, Friedrich JC, O'Dea RE, Holcombe AO. 2024 Team scientists should normalize disagreement. *Science* **384**, 1076–1077. (doi:10.1126/science.ado7070)

64. Aczel B *et al*. 2021 Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife* **10**, e72185. (doi:10.7554/elife.72185)

65. Obels P, Lakens D, Coles NA, Gottfried J. 2020 Analysis of open data and computational reproducibility in registered reports in psychology. *Adv. Methods Pract. Psychol. Sci.* **3**, 229–237. (doi:10.1177/2515245920918872)

66. Hardwicke TE, Vazire S. 2023 Transparency is now the default at *Psychological science*. *Psychol. Sci.* **35**, 708–711. (doi:10.1177/09567976231221573)

67. Cracking the code review process. *Nat. Comput. Sci.* **2**, 277 (2022). (doi:10.1038/s43588-022-00261-w)

68. Ivimey-Cook ER *et al*. 2023 Implementing code review in the scientific workflow: insights from ecology and evolutionary biology. *J. Evol. Biol.* **36**, 1347–1356. (doi:10.1111/jeb.14230)

69. Breznau N *et al*. 2022 Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proc. Natl Acad. Sci. USA* **119**, e2203150119. (doi:10.1073/pnas.2203150119)

70. Lohr S. 2020 Universities and tech giants back national cloud computing project. *New York Times*. See https://www.nytimes.com/2020/06/30/technology/national-cloud-computing-project.html.

71. Coles NA, Perz B, Behnke M, Eichstaedt JC, Kim SH, Vu TN. 2025 Data from: 2023 Emotion physiology and experience collaboration (EPiC) challenge. *Open Science Framework*. (doi:10.17605/OSF.IO/BMHSD)

72. Coles N, Perz B, Behnke M, Eichstaedt J, Kim SH, Vu T *et al*. 2025 Supplementary material from: Big team science reveals promises and limitations of machine learning efforts to model physiological markers of affective experience. Figshare. (doi:10.6084/m9.figshare.c.7828922)