nature research

Corresponding author(s): Alan S. Cowen

Last updated by author(s): Oct 2, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Custom and proprietary code was used to search for videos and process the video content and metadata using machine learning algorithms. These data collection and processing steps can partially be replicated using the Google Cloud Video Intelligence and Natural Language APIs, as referenced in the manuscript. Anonymized (differentially private) versions of the context-expression correlations in each country will be made available via Github under the repository github.com/alanscowen/contextexpression.
Data analysis	Analysis of the processed data was performed using custom code in Matlab version R2018B. Code to read and visualize the anonymized context-expression correlations in each country will be made available via Github under the repository github.com/alanscowen/contextexpression.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data presented in the manuscript was publicly available at the time of data analysis and largely remains publicly available, although the data is owned by the original YouTube contributors who are free to remove their videos from the Internet any time. However, we are unable to release identifiers of the specific videos we analyzed. Anonymized (differentially private) versions of the context-expression correlations in each country for each experiment are available in github.com/alanscowen/contextexpression. The MIT CBCL Database (used in Extended Data Figure 2) is available upon request at http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A quantitative observational study correlating facial expressions with other video content in YouTube videos.
Research sample	A total of approximately 6 million videos uploaded to YouTube. Experiment 1: We sought to focus on natural footage for which reliable geographic information was available. To find naturalistic footage, we searched for publicly available YouTube videos that were uploaded from mobile phones. The search was restricted to YouTube videos tagged with a latitude and longitude of upload that matched the country in which the uploader was registered. Furthermore, to focus on naturalistic footage, we filtered out videos predicted by the video topic annotations to include video games and other animated content. This yielded a total of 3,029,812 videos. Experiment 2: The videos from Experiment 1, like many videos on YouTube, typically lacked detailed descriptions, making them poor candidates for annotation by the text topic DNN. Thus, we collected a new set of videos for Experiment 2. To ensure that we would have the power to investigate correlations between contexts and facial expressions, we sought to include publicly available videos that had titles and descriptions pertaining either to the contexts we explored in Experiment 1 or to emotions. To do so, we first searched for videos with a wide range of context- and emotion-related substrings within their English-translated titles and descriptions (Dataset S3; note that to the extent that translations were inaccurate, representation of corresponding contexts could be reduced in non-English-speaking cultures, exacerbating cultural differences). We then retrieved the full native-language titles and descriptions for those videos and computed text topic annotations. Finally, to avoid synthetic faces, we filtered out videos predicted by the text topic DNN to include video games and animated content. This yielded a total of 3,056,861 videos.
Sampling strategy	N/A -- Full population of videos meeting the criteria specified above were included in the study.
Data collection	Processing of YouTube videos was performed on temporary cloud computing system instances without permanently downloading any video data or metadata.
Timing	Videos included in Experiment 1 were uploaded between July 14, 2009 and May 3, 2018. Videos included in Experiment 2 were uploaded between December 27, 2005 and April 15, 2019. Facial expression annotations were generated between May 3, 2018 and May 1, 2019. Context annotations were generated between the time of upload of each video and May 1, 2019. All statistical analyses were performed between May 3, 2018 and May 1, 2019.
Data exclusions	All data meeting the criteria specified above were included in the study.
Non-participation	N/A
Randomization	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	N/A
Ethics oversight	The use of the video data in aggregate form underwent review for alignment with Google's AI Principles (see https://ai.google/principles/) and conformed to Google's privacy policy (see https://policies.google.com/privacy).

Note that full information on the approval of the study protocol must also be provided in the manuscript.