PLOS BIOLOGY

PERSPECTIVE

Neuroimaging-based prediction of mental traits: Road to utopia or Orwell?

Simon B. Eickhoff[1,2]*, **Robert Langner**[1,2]*

1 Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, **2** Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Centre Jülich, Jülich, Germany

* s.eickhoff@fz-juelich.de (SBE); robert.langner@hhu.de (RL)

Abstract

Predicting individual mental traits and behavioral dispositions from brain imaging data through machine-learning approaches is becoming a rapidly evolving field in neuroscience. Beyond scientific and clinical applications, such approaches also hold the potential to gain substantial influence in fields such as human resource management, education, or criminal law. Although several challenges render real-life applications of such tools difficult, future conflicts of individual, economic, and public interests are preprogrammed, given the prospect of improved personalized predictions across many domains. In this Perspective paper, we thus argue for the need to engage in a discussion on the ethical, legal, and societal implications of the emergent possibilities for brain-based predictions and outline some of the aspects for this discourse.

Citation: Eickhoff SB, Langner R (2019) Neuroimaging-based prediction of mental traits: Road to utopia or Orwell? PLoS Biol 17(11): e3000497. https://doi.org/10.1371/journal.pbio.3000497

Published: November 14, 2019

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Provenance: Not commissioned; externally peer reviewed.

Many potentially life-altering decisions that are made about a person by someone else involve judgments about "internal" characteristics like intelligence, trustworthiness, or other mental traits, that is, aspects that are not directly observable for the judge. This makes judgments difficult and error prone. For example, a company may want to hire a manager who is strongly determined and also very open to collaboration. Naturally, all applicants assert that they have these traits, so whom to select? As another example, a judge needs to decide whether counseling during incarceration has reduced aggressive tendencies to a level that does not pose a risk for others.

Traditionally, such questions have been tackled by extended interviews and looking at potential discrepancies between self-reported characteristics and previous behavior. This not only limits objectivity because of examiner effects but will also be biased by the degree the interviewees can "sell themselves" (i.e., their impression management skills), curtailing the validity of such assessments. Recent advances in the application of machine learning and artificial intelligence (AI) toward the predictive analysis of brain imaging data, however, may induce a disruptive change of this situation. Several studies now suggest that not only age or gender but also complex mental traits such as intelligence [1], attentional abilities [2], altruism [3], or personality factors [4] may be predicted in individuals from brain imaging data.

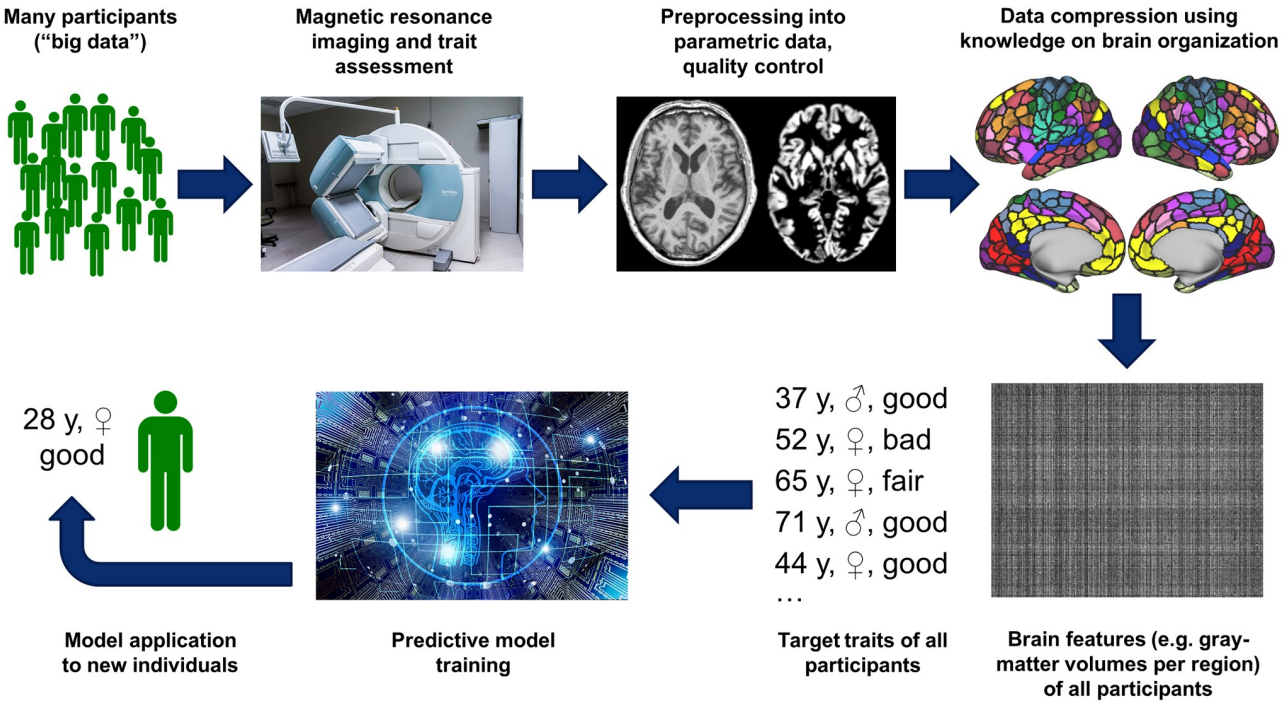

Fig 1. Schematic sketch of a pipeline for building brain-based prediction models for individual traits. To be read clockwise starting at the top left. *Parcellated brain hemispheres (top right panel) reproduced from* [7] *under a CC BY 4.0 license* (https://creativecommons.org/licenses/by/4.0/), *depicting results reported in* [8].

https://doi.org/10.1371/journal.pbio.3000497.g001

Notwithstanding heterogeneity of technical implementations (see the aforementioned papers and [5,6]), the approach can be summarized as follows: Structural or functional (resting-state) neuroimaging data as well as the target trait measure are collected in large samples comprising several hundred participants. After preprocessing and a representation of individual neurobiology as parametric values, a machine-learning model is trained to find a mapping from the imaged brain features to the trait of interest. Generalization of the model is then assessed by predicting that trait in previously unseen people, either in an independent sample or through cross validation, and comparing the predicted with the (known) actual phenotype (Fig 1).

Have we as cognitive neuroscientists thus provided the basis for more objective and valid assessments of personal aptitudes, attitudes, and other mental characteristics, making everyone's life better? If these methods are further developed and widely adopted, the entire society might benefit in many important aspects from improved evaluation and decision procedures that are devoid of implicit biases such as halo effects or other judgmental errors on the part of the observer. Besides being fairer and supporting equal opportunities for all, more valid assessments would engender more accurate matches between personal characteristics and contextual factors (e.g., specific therapies, job demands, or stressors), potentially enhancing health, life satisfaction, and productivity (compare with [9]). Or have we opened Pandora's box and paved the way for an Orwellian future in which algorithms "know" our innermost features and dictate our potential choices in life? Ultimately, these questions can only be answered in the retrospective.

Although there has been increasing interest in the ethical and legal implications of neuroscientific progress since the early 2000s [10–15], neuroimaging-based prediction has only recently advanced to a degree that put it on the map for neuroethical discourse ([16]; see [17] for a review on using neuroimaging for violence prediction in legal contexts). At present, it

seems appropriate to take a realistic look at the potential and limitations of these techniques and to identify issues for societal discussion.

First, a distinction must be made between scientific demonstrations of predictive power (e.g., a significant correlation between true and predicted traits in new participants) and algorithms that can be successfully used in real-life diagnostic or prognostic settings. To illustrate the point, a correlation of r = .70 would be considered a strong effect for group-level associations and, in fact, is probably the best that can currently be achieved for complex traits. However, it still explains only about half of the variance in the target variable. This exemplifies the frequently observed discrepancy between "statistically significant" and "practically relevant." It should be noted, though, that predictive models in neuroimaging are not only developed for personalized predictions, as focused on in this Perspective, but also with another goal in mind: to identify generalizable brain–behavior relationships. And for this purpose, finding substantial statistical associations like r = .70 would be considered highly relevant.

But how precise must an algorithm be to become relevant in applied settings entailing individual assessment? The answer obviously relates to the severity of the consequences of erroneous predictions. First of all, if a certain characteristic is rare, even a highly precise algorithm will produce many misclassifications and associated adverse consequences when used in large-scale evaluations (e.g., 90% accuracy will yield 100 errors in 1,000 cases screened). That said, would we accept 90% prediction accuracy in the context of a hiring decision? Most likely. But would it be acceptable for releasing an apparently rehabilitated child molester from detention? The majority answer might rather be "no" in this case.

In these scenarios, human evaluators also make mistakes and might barely fare better than (hypothetical) algorithms. Hence, do we impose higher demands for accuracy on AI? It seems so, but should this be the case? First of all, by using AI support, we aim to improve predictions and decision processes beyond the current human standards. Another part of the answer to this question, however, may be a lack of trust in AI because of its lack of discursive capacity: humans can present their thought processes and conclusions—even if partly confabulated post hoc because of the limits of introspection—which, in turn, allows others to integrate the decision with their own experience and knowledge and emulate and appraise the decision process. Algorithm-based predictions usually lack this (potentially spurious) explainability, which may constitute an obstacle to their broader societal acceptance. In addition, making life-impacting decisions might feel strange and discomforting or even illegitimate to many if it were solely driven by machine output, even if AI-based predictions were somewhat more accurate than human-made ones. How to weight human traceability and other "soft" features of the decision process vis-à-vis verifiably precise but unfathomable "black boxes" will most likely depend on the degree to which AI-supported algorithms reliably outperform human decision-makers.

Second, we need to acknowledge that the brain is not static and there is no one-way road from brain to mind (i.e., there is no unidirectional causal one-to-one mapping from brain activity to mental phenomena). Hence, we as human beings are not subject to a predefined fate coded in our neurobiology. This is particularly true when it comes to longer-term predictions, which may be of particular interest in many applications. Given the plasticity of the human brain, both the effects of agency (e.g., voluntary changes in lifestyle or approach) and outside influences may substantially impact the behavioral outcome of interest as well as the brain itself. For example, a job candidate may be predicted to be not well suited for a particular task but successfully works on herself to adapt to the challenges of the job, rendering the prediction invalid. Conversely, a criminal offender may have responded well to treatment and gets a very favorable prediction yet reverts to a problematic lifestyle after returning into previous social settings. How to accommodate such widening of the "predictive funnel" with time (i.e., the growing imprecision with increasing temporal distance to the predicted event) in

neuroimaging-based predictions of behavior remains an open issue. This is also true for weather or traffic jam forecasts, to name just two examples of yet-unsolved prediction difficulties in complex dynamical systems in which the basic physical laws ruling the interactions of different factors are known—something that cannot be said of structure–function relations in the brain, let alone brain–behavior relationships. Noting that similar considerations hold for current expert assessments, we would argue that brain-based predictions should stimulate the respective discussions through quantitative estimates of predictive funnels.

Besides, the growing imprecision for temporally more distant events might be ameliorated by moving away from the binary nature of many prognoses (e.g., responsive versus not responsive to a given training, suitable versus not suitable for a particular job, or given versus not given to violence) toward time-sensitive continuous risk models as proposed by Matthew Baum [16]. This kind of probabilistic modeling has already been successfully adopted in other domains, such as forecasting rain and other weather conditions. Further, to accommodate the impact of contextual (nonbrain) factors like particular behaviors or social and environmental settings, pertinent data from smartphones and other wearable devices could provide complementary information to enrich and improve "neuroprediction" models.

Third, an oft-underestimated aspect in projections of future use is the discrepancy between technical and practical feasibility. The resources needed to assess hundreds or, more likely, thousands of people using neuroimaging are substantial, particularly when following these people longitudinally over months or years to observe a relevant (future) outcome. Furthermore, building practically relevant prediction models will likely require rather extensive imaging from each participant to achieve sufficient reliability despite the brain's nonstationarity and potentially also multimodal data to cover various relevant aspects of neurobiology. For all this, highly cooperative participants are needed, also to achieve an appropriate level of data quality, as neuroimaging data are notoriously noisy and easily distorted or ruined by noncompliant behavior during scanning. Taken together, this is a huge challenge for developing as well as applying such models in real life, as the best model can only work if it gets all the input it requires. Last but not least, all these efforts will be futile if the quality of behavioral (psychometric) trait assessment is all but very high, as brain–behavior associations can never be closer—and thus, brain-based trait predictions never more precise—than is the level of reliability on either side ([18]; see also [19]). We need to keep in mind that traditional assessment procedures, although being the "gold standard" of trait measurement against which new prediction algorithms are evaluated, do not reveal the ground truth but come with their own shortcomings, as alluded to before.

Given these difficulties, how realistic are the promises and expectations outlined previously? In fact, the current picture is mixed: the initial prediction successes, which were too limited for real-life use, could not be markedly improved on by using larger samples ([20]; but see [21]). Also, even complex multivariate assessments like connectomic fingerprinting seem to be less individual and robust than expected [22]. This state of affairs likely results from a mixture of the aforementioned difficulties and other issues, some of which are beginning to be addressed, such as large-scale multimodal imaging and modeling. Furthermore, new markers of brain function and connectivity are likely to be identified, and prediction methods are going to be improved.

At any rate, as neuroimaging is rather costly, relative to other established or novel methods that may yield potentially predictive biomarkers (e.g., smartphones, ambulatory assessments, or electroencephalography), prediction based on neuroimaging data must be shown to clearly outperform competing approaches to justify its costs. From today's perspective, given the remaining challenges, it seems unlikely that this kind of neuroprediction of mental traits will ever be universally applicable. A realistic expectation, though, might be its practical application

in certain fields for specific questions, particularly when important issues are at stake for which other valid prognostic information is not available or otherwise obtainable.

The bottom line is that highly precise imaging-based prediction of mental traits in real-world scenarios requires substantial investments. Without these efforts, the ultimate potential of the outlined methods remains theoretical. It goes without saying that such challenges have far better chances to be met in settings with strong commercial or political interests of financially potent players. In such scenarios, however, conflicts of interest become an integral part of the process, and questions on permissibility arise. For instance, should an insurance company or a hospital group be allowed to train models on the data of their clients to predict future illness, even after obtaining individual consent to such data usage by their clients? Hardly anyone would disagree when the goal is to improve preventive care. But what if exactly the same data and results are used to cancel insurance coverage?

This illustrates an ethical issue previously discussed in regard to genetic data: the potential proliferation of inferential opportunities (compare with [23]). Data gained from conducting interviews, psychometric testing, or administering self-report inventories can mostly serve only the purpose it was collected for, whereas neuroimaging (like genetic) data, once collected, could be successfully submitted to a much broader number of predictive models, including those that were not yet thinkable when the data were acquired. Acknowledging the aforementioned aspects of plasticity, a brain scan obtained for an unrelated medical purpose could later be reused to assess, say, tendencies to violence and political extremism. Although this example is yet purely fictional, it still illustrates the potential uncontrollable misuse of brain imaging data. Considering how readily people are sharing genetic data with commercial companies, such a scenario could lead to a flourishing secondary market for predictive material. This obviously also applies to behavioral data, including verbal communication, obtained from mobile devices like smartphones and other wearables because of the broad scope of such data and the continuity of their collection, especially when combined with neuroimaging data as mentioned previously. Such considerations evidently lead to questions of data ownership, including the right to have data deleted, the limits of informed consent, as well as the weighting of personal and public interests. If and how neuroimaging data that could disclose personal information may be analyzed by current or future prediction algorithms is a question that only will grow in relevance when considering that through advanced data analysis, more and more types of data may yield predictive personal information in the future.

To conclude, it depends on us whether advances in the neuroimaging-based prediction of mental traits will move us closer to some form of utopia or drive us toward some Orwellian dystopia. Even if still a long, obstacle-strewn road ahead in any case, the core ethical and legal issues should be addressed now to avoid undesirable facts being established by individual stakeholders.

References

1. Dubois J, Galdi P, Paul LK, Adolphs R. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. Philos Trans R Soc Lond B Biol Sci. 2018; 373(1756). https://doi.org/10.1098/rstb.2017.0284 PMID: 30104429
2. Rosenberg MD, Finn ES, Scheinost D, Papademetris X, Shen X, Constable RT, et al. A neuromarker of sustained attention from whole-brain functional connectivity. Nat Neurosci. 2016; 19(1): 165–71. https://doi.org/10.1038/nn.4179 PMID: 26595653
3. Tusche A, Bockler A, Kanske P, Trautwein FM, Singer T. Decoding the charitable brain: Empathy, perspective taking, and attention shifts differentially predict altruistic giving. J Neurosci. 2016; 36(17): 4719–32. https://doi.org/10.1523/JNEUROSCI.3392-15.2016 PMID: 27122031
4. Nostro AD, Muller VI, Varikuti DP, Plaschke RN, Hoffstaedter F, Langner R, et al. Predicting personality from network-based resting-state functional connectivity. Brain Struct Funct. 2018; 223(6): 2699–719. https://doi.org/10.1007/s00429-018-1651-z PMID: 29572625

5. Scheinost D, Noble S, Horien C, Greene AS, Lake EM, Salehi M, et al. Ten simple rules for predictive modeling of individual differences in neuroimaging. NeuroImage. 2019; 193: 35–45. https://doi.org/10.1016/j.neuroimage.2019.02.057 PMID: 30831310
6. Shen X, Finn ES, Scheinost D, Rosenberg MD, Chun MM, Papademetris X, et al. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nat Protoc. 2017; 12(3): 506–18. https://doi.org/10.1038/nprot.2016.178 PMID: 28182017
7. Kong R. Schaefer2018_400regions_17networks [figure]; 2019 [cited 2019 Oct 28]. Database: figshare [Internet]. https://figshare.com/articles/Schaefer2018_400regions_17networks/10059701.
8. Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb Cortex. 2018; 28(9): 3095–114. https://doi.org/10.1093/cercor/bhx179 PMID: 28981612
9. Gabrieli JDE, Ghosh SS, Whitfield-Gabrieli S. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron. 2015; 85(1): 11–26. https://doi.org/10.1016/j.neuron.2014.10.047 PMID: 25569345
10. Farah MJ. Emerging ethical issues in neuroscience. Nat Neurosci. 2002; 5(11): 1123–9. https://doi.org/10.1038/nn1102-1123 PMID: 12404006
11. Farah MJ. Neuroethics: the practical and the philosophical. Trends Cogn Sci. 2005; 9(1): 34–40. https://doi.org/10.1016/j.tics.2004.12.001 PMID: 15639439
12. Gazzaniga MS. The ethical brain. Washington, DC: Dana Press; 2005.
13. Illes J, Sahakian BJ, editors. Oxford handbook of neuroethics. New York: Oxford University Press; 2011.
14. Chatterjee A, Farah MJ, editors. Neuroethics in practice. New York: Oxford University Press; 2013.
15. Greely HT, Ramos KM, Grady C. Neuroethics in the age of brain projects. Neuron. 2016; 92(3): 637–41. https://doi.org/10.1016/j.neuron.2016.10.048 PMID: 27810008
16. Baum ML. The neuroethics of biomarkers: What the development of bioprediction means for moral responsibility, justice, and the nature of mental disorder. New York: Oxford University Press; 2016.
17. Poldrack RA, Monahan J, Imrey PB, Reyna V, Raichle ME, Faigman D, et al. Predicting violent behavior: What can neuroscience add? Trends Cogn Sci. 2018; 22(2): 111–23. https://doi.org/10.1016/j.tics.2017.11.003 PMID: 29183655
18. Vul E, Harris C, Winkielman P, Pashler H. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect Psychol Sci. 2009; 4(3): 274–90. https://doi.org/10.1111/j.1745-6924.2009.01125.x PMID: 26158964
19. Gignac GE, Bates TC. Brain volume and intelligence: The moderating role of intelligence measurement quality. Intelligence. 2017; 64: 18–29. https://doi.org/10.1016/j.intell.2017.06.004
20. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci. 2017; 20(3): 365–77. https://doi.org/10.1038/nn.4478 PMID: 28230847
21. Cui Z, Gong G. The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. NeuroImage. 2018; 178: 622–37. https://doi.org/10.1016/j.neuroimage.2018.06.001 PMID: 29870817
22. Waller L, Walter H, Kruschwitz JD, Reuter L, Muller S, Erk S, et al. Evaluating the replicability, specificity, and generalizability of connectome fingerprints. NeuroImage. 2017; 158: 371–7. https://doi.org/10.1016/j.neuroimage.2017.07.016 PMID: 28710040
23. Tavani HT. Genomic research and data-mining technology: implications for personal privacy and informed consent. Ethics Inf Technol. 2004; 6(1): 15–28. https://doi.org/10.1023/b:etin.0000036156.77169.31 PMID: 16969958