Decoding mood

Amit Etkin

Invasive recordings of human brain activity predict mood fluctuations.

The subjective, shifting nature of mood has long proved recalcitrant to brain-imaging technologies. In this issue, Sani *et al.*[1] overcome this roadblock for the first time with a report on a machine-learning model capable of predicting mood using only electrical signals from the brain. Their findings open an exciting new door into the possibilities of decoding mood-related neural signals in humans.

For many years, recordings of brain activity have been used to decode hidden internal states, but this research has focused almost exclusively on motor control. In a typical experiment, machine learning is used to associate signals from invasive or noninvasive neural recordings with an organism's intention to move a part of the body. Models trained in this way can predict a person's intended movements from measured brain activity alone, and their accuracy is determined by how well the algorithm recapitulates the observed movements. This work has led to rapid advances in brain/machine-interface technology and in understanding how the brain encodes motor control[2,3]. However, the question of whether a similar approach could be used to decode mood—a more complex and ill-defined state—has remained unanswered.

Mood is particularly challenging because it is a hard-to-define psychological construct. There is no ground truth against which computational models learned from brain recordings can be reliably indexed, and the only access to mood is through individual self-reporting by questionnaire-guided introspection. In addition, although mood can vary across time scales of minutes to months, mood measurements cannot provide the high-precision measurements possible in motor-control decoding.

Amit Etkin is at the Department of Psychiatry and Behavioral Sciences and Stanford Neurosciences Institute, Stanford University, Stanford, California, USA, and at the Veterans Affairs Palo Alto Healthcare System and the Sierra Pacific Mental Illness, Research, Education, and Clinical Center (MIRECC), Palo Alto, California, USA.
e-mail: *amitetkin@stanford.edu*

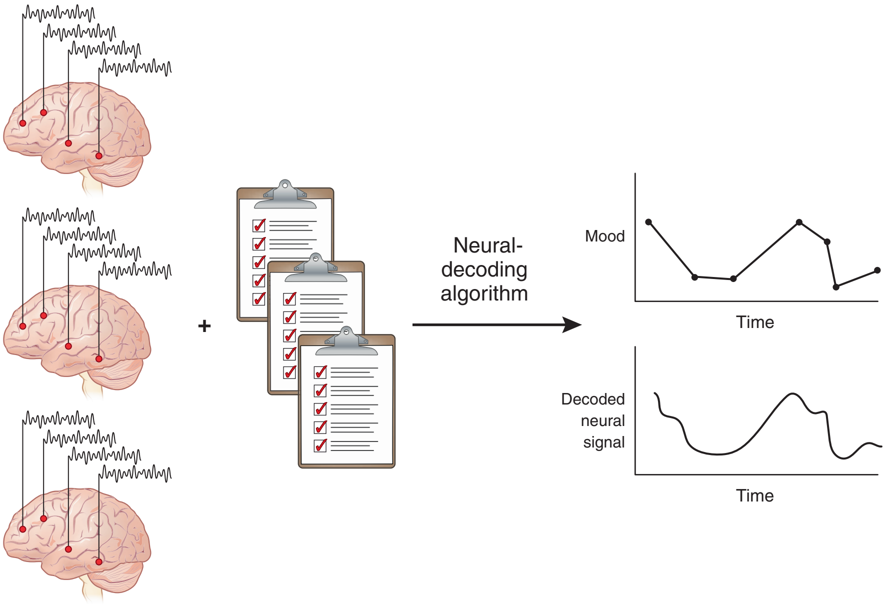

Figure 1 Experimental design and analysis approach. Multifocal invasive neural recordings at multiple time points in humans were aligned with the administration of questionnaire-based mood measurements and used to decode neural signals that tracked hourly mood fluctuations in the participants.

In their new study, Sani *et al.*[1] started by measuring invasively recorded electrocorticography (ECoG) signals from multiple regions across the brains of patients admitted to the hospital for monitoring of their epilepsy. Typically, information from such recordings is used to localize epileptic foci to guide seizure-reducing interventions, such as tissue resection or implantation of neurostimulation devices. The authors built on extensive literature assessing momentary changes in mood and extended these approaches to patients undergoing ECoG recordings. Patients filled out a questionnaire designed to capture their short-term mood fluctuations every ~12 hours for 6 days of continuous ECoG recordings (**Fig. 1**). Recording electrodes were implanted in regions such as the amygdala, hippocampus, orbitofrontal cortex and anterior cingulate cortex—limbic and paralimbic regions that are known to respond to a variety of emotional cues—as well as a limited selection of nonlimbic regions.

The authors developed a mood-decoding machine-learning model for each of the seven study participants. The models were trained on questionnaire data and ECoG recordings from each individual to identify variations in the neural data that corresponded to changes in the subject's mood. To validate the approach, the authors performed leave-one-out cross-validation, in which the reported mood at one time point was omitted from the training set, and the remaining data were used to predict the missing measurement. This procedure was repeated for all time points. Model performance was assessed by calculation of the error between the predicted and observed values, and the models succeeded to various degrees, explaining 12–65% of the mood fluctuation.

Given the large amount of recorded neural data and the relatively limited frequency of mood sampling, a critical element to the success of this study was the nature of the mood-decoding machine-learning algorithm. The model is an example of latent-space models, which are increasingly being used in neuroscience[4,5]. In such models, which may take many forms, the data are reduced to capture the most important prediction-relevant low-dimensional representations of high-dimensional and potentially very noisy raw signal. The authors' success in decoding mood was thus attributable both to their novel use of repeated mood measurements in epilepsy patients and to a decoding model optimized for parsimony given the sparse nature of the mood data.

Several important findings emerge from this work. First, decoding was found to rely centrally on limbic/paralimbic regions rather than regions outside the limbic system, confirming previous knowledge about the effects of these regions on mood. Second, the decoder models were found to be stable across time points for the individual. Moreover, once decoded, the neural signals corresponding to their model of mood fluctuations could be tracked throughout the recording period, yielding a potential ongoing neural correlate of mood (**Fig. 1**). However, the same analyses on neural data drawn from periods outside those corresponding to when participants filled out the questionnaires did not yield successful decoding. Thus, the neural signals might indicate mood only when measured while individuals are engaged in filling out questionnaires (i.e., might be state dependent), or the brain–mood relationship may be very temporally precise, probably because of rapid fluctuations in either the mood or neural signals. The results may also reflect overfitting of the predictive models. Overfitting is a particular concern, given the authors' use of leave-one-out cross-validation (which inflates model accuracy) and the limited number of mood samples per patient. Future replication will be needed to understand whether overfitting occurred.

Although the work of Sani *et al.*[1] is highly innovative, it raises several points for further consideration. The clinical relevance of the observed hourly fluctuations in mood is unknown, and the relationship of these fluctuations to mood-related psychopathology is likewise unclear. Psychology distinguishes between emotions, which are transient and typically stimulus triggered, and moods, which are more prolonged and are driven by external cues and internal states. Where exactly the questionnaire used in this study falls in measuring emotions or mood is uncertain. Moreover, on the aggregate, the authors' model explained only 32% of the variability in mood across the 6 days of recordings, leaving most of the mood-related signal in the questionnaires unexplained.

The authors note that decoding in all individuals involved signals from limbic/paralimbic regions, and not regions outside this network. However, the specific brain regions responsible for the mood-decoding signals differed sharply between individuals, even within the broadly defined limbic network. For example, the orbitofrontal cortex was present in only four of seven individuals. The hippocampus and dorsal cingulate regions were present in only two of seven individuals. Given that each of these regions engages in very different neural computations, the degree of interindividual variability in mood-decoding regions was striking. Thus, an optimal neural mood 'signature' may have to be decoded separately for each person. The extent to which this specificity is real versus due to model overfitting is also difficult to know at this point, but should become more clear in replication studies.

Although the nonlimbic regions examined by the authors did not support decoding, the authors studied very few such regions (primarily the temporal cortex). Given the broad cortical interconnections of the limbic system and that cortical regions in this system contributed more heavily to decoding than deep brain regions, an exciting question is whether conventional scalp electroencephalography recordings might be used to decode mood in a manner similar to the invasive recordings used here.

From a therapeutic perspective, understanding how neural signals change in response to mood may lead to new treatments for neuropsychiatric disorders. If the approach of Sani *et al.*[1] can be generalized to other recording and mood-measurement methods, one might imagine, for example, that such signals could provide the afferent limb to a closed-loop neurostimulation approach for modifying mood. In such a scenario, after the mood states of patients with a condition such as depression had been decoded, signals that trigger a stable and more positive mood state could be provided. Similar decoding of other subjective experiences could lead to a better understanding of other emotions, including those central to both psychiatric conditions and normative emotional experience. Such work could help decipher the uniqueness of subjective experience and whether its neural correlates differ as a function of factors such as a psychiatric diagnosis, certain medications or even age.

Although initial insights in any domain of investigation often raise more questions than they answer, as is the case here, the authors deserve to be congratulated on path-breaking work that will no doubt inspire many others in the field.

COMPETING INTERESTS

A.E. has received stock options in Mindstrong Health and Akili Interactive for unrelated consulting.

1. Sani, O.G. *et al. Nat. Biotechnol.* **36**, 954–961 (2018).
2. Brandman, D.M., Cash, S.S. & Hochberg, L.R. *IEEE Trans. Neural Syst. Rehabil. Eng.* **25**, 1687–1696 (2017).
3. Li, Z. *Front. Syst. Neurosci.* **8**, 129 (2014).
4. Kato, S. *et al. Cell* **163**, 656–669 (2015).
5. Williams, A. H. *et al. Neuron* **98**, 1099–1115.e1098 (2018).