Predicting the knowledge–recklessness distinction in the human brain

Iris Vilares[a,b,1], Michael J. Wesley[c,1], Woo-Young Ahn[d], Richard J. Bonnie[e], Morris Hoffman[f], Owen D. Jones[g,h], Stephen J. Morse[i], Gideon Yaffe[j,2], Terry Lohrenz[b], and P. Read Montague[a,b,2]

[a]Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, United Kingdom; [b]Virginia Tech Carilion Research Institute, Virginia Tech, Roanoke, VA 24016; [c]Department of Behavioral Science, University of Kentucky College of Medicine, Lexington, KY 40506; [d]Department of Psychology, Ohio State University, Columbus, OH 43210; [e]Institute of Law, Psychiatry and Public Policy, University of Virginia, Charlottesville, VA 22903; [f]Second Judicial District (Denver), State of Colorado, Denver, CO 80202; [g]Vanderbilt Law School, Vanderbilt University, Nashville, TN 37203; [h]Department of Biological Sciences, Vanderbilt University, Nashville, TN 37203; [i]University of Pennsylvania Law School, University of Pennsylvania, Philadelphia, PA 19104; and [j]Yale Law School, Yale University, New Haven, CT 06511

Edited by Terrence J. Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, and approved February 9, 2017 (received for review November 23, 2016)

Criminal convictions require proof that a prohibited act was performed in a statutorily specified mental state. Different legal consequences, including greater punishments, are mandated for those who act in a state of knowledge, compared with a state of recklessness. Existing research, however, suggests people have trouble classifying defendants as knowing, rather than reckless, even when instructed on the relevant legal criteria. We used a machine-learning technique on brain imaging data to predict, with high accuracy, which mental state our participants were in. This predictive ability depended on both the magnitude of the risks and the amount of information about those risks possessed by the participants. Our results provide neural evidence of a detectable difference in the mental state of knowledge in contrast to recklessness and suggest, as a proof of principle, the possibility of inferring from brain data in which legally relevant category a person belongs. Some potential legal implications of this result are discussed.

neurolaw | mental states | knowledge | recklessness | elastic-net model

Imagine you are a juror in the trial of a defendant who admits to having transported a suitcase full of drugs across international borders. However, you do not know how aware she was of the presence of drugs in that suitcase. The degree of awareness she had at the time she crossed the border will make a difference to her criminal culpability and, in turn, to the amount of punishment she faces.

Conviction for a crime requires proof beyond a reasonable doubt of both the crime's *actus reus*—a set of statutorily specified acts, results, and circumstances, such as crossing a border while in possession of drugs—and the crime's *mens rea*—a set of statutorily specified mental states including, for instance, knowledge that one is in possession of drugs when one crosses the border. The Model Penal Code (MPC), which is followed in many jurisdictions in the United States, distinguishes among four different psychological states a person can be in with respect to each element of a crime's *actus reus*: purpose, knowledge, recklessness, and negligence. The Code also specifies that these decrease in culpability: it is worse, for instance, to cross the border knowing you have drugs (as one is if sure that one has them) than to do so while reckless with respect to that fact (as one is if aware of a "substantial and unjustifiable risk" that one is carrying drugs, but uncertain that one is) (MPC §2.02). The MPC's four-part taxonomy, however, relies on at least two assumptions: (*i*) people actually differ psychologically in the ways that the MPC sets out; and (*ii*) average people (potential jurors) can effectively categorize real-world mental states in accordance with the Code's definitions (1). Considering the dramatic effects that different mental-state assignments can have on the freedom of criminal defendants, it is surprising that very little research has been done to verify these assumptions (1, 2).

Shen et al. (1), setting out to test the second assumption, recruited participants from different parts of the United States, gave them different crime scenarios, and asked them to identify which of the four mental states the protagonist of the scenario was in. The research revealed that, although people were quite good at distinguishing between intentional, negligent, and blameless (no culpability) states, their ability to distinguish between a knowing and a reckless state was surprisingly poor, with people confusing the two about 45% of the time. Nevertheless, in a real court, to judge someone to have knowingly rather than recklessly committed a criminal act can make an enormous difference in punishment. In fact, it can be, literally, a matter of life and death: a defendant can be eligible for the death penalty if found to have performed a lethal act knowing it would kill rather than merely aware of a substantial risk that it would. With an individual's freedom and potentially life hanging in the balance, it seems necessary to find multiple and reliable ways to facilitate accurate sorting between knowing and reckless mental states. To this end, scientific evidence for (or against) biologically based and brain-based distinctions of knowing and reckless mental states, and the boundary that may separate them, could help us either to refine or to reform the ways criminal responsibility is assessed.

Currently, the most frequently used tool to study the neural correlates of "mental states" is functional magnetic resonance imaging (fMRI) (3). fMRI analysis has been recently used in the context of the law, from trying to predict psychopathy (4) to trying to understand what goes on in the brains of jurors when

Significance

Because criminal statutes demand it, juries often must assess criminal intent by determining which of two legally defined mental states a defendant was in when committing a crime. For instance, did the defendant know he was carrying drugs, or was he merely aware of a risk that he was? Legal scholars have debated whether that conceptual distinction, drawn by law, mapped meaningfully onto any psychological reality. This study uses neuroimaging and machine-learning techniques to reveal different brain activities correlated with these two mental states. Moreover, the study provides a proof of principle that brain imaging can determine, with high accuracy, on which side of a legally defined boundary a person's mental state lies.

Author contributions: R.J.B., M.H., O.D.J., S.J.M., G.Y., T.L., and P.R.M. designed research; I.V. and M.J.W. performed research; W.-Y.A. contributed new reagents/analytic tools; I.V., M.J.W., T.L., and P.R.M. analyzed data; and I.V., M.J.W., W.-Y.A., R.J.B., M.H., O.D.J., S.J.M., G.Y., T.L., and P.R.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]I.V. and M.J.W. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: read@vtc.vt.edu or gideon.yaffe@yale.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1619385114/-/DCSupplemental.

they are deciding whether to punish (5). However, no fMRI studies of which we are aware have attempted to determine whether and how the "culpable mental states," as defined by the MPC, map onto differential activations in the human brain.

Given that the main distinction between the knowing and reckless mental states relies on the differential perception of probabilities and uncertainty associated with an outcome (if knowing you are "practically certain" of the outcome, i.e., $P = 1$, whereas if reckless you are aware of a "substantial" risk but uncertain, i.e., $0 < P < 1$), potential brain areas differentially associated with the knowing or reckless mental states could be areas previously found in the neuroeconomics and decision-making literature to be implicated in encoding probability or uncertainty and risk (6–11). These areas include the posterior parietal cortex (7, 12), the posterior cingulate cortex (12, 13), the medial and lateral prefrontal cortex (6, 12), the thalamus (7, 8), and the insula (9, 10). However, these studies almost always use simple lotteries or gambling tasks (e.g., choice between two decks of cards; guessing from which urn a ball came from) and do not portray a legally relevant knowing vs. reckless situation.

Although typical fMRI analyses are descriptive in nature and lack predictive power, new methods are emerging that try to find multiregional brain activity patterns that collectively predict a specific cognitive condition or individual characteristic (14–20). This is a particularly challenging task, given that, with fMRI data, the number of predicting variables is generally much higher than the number of observations, and hence there is a risk of producing either computationally intractable or strongly overfit models (15, 21, 22). A new method has been suggested that tries to tackle this problem by using elastic-net (EN) regression. EN regression uses a mix of L1 and L2 regularization to prevent overfitting, while at the same time ensuring that the final model includes all of the relevant brain regions (14, 15, 21). This new method could potentially be applied to predict the MPC's "culpable" mental states based on a person's fMRI data.

In this study, we attempt to understand whether knowledge and recklessness are actually associated with different brain states, and which are the specific brain areas involved. Moreover, we want to know whether it is possible to predict, based on brain-imaging data alone (using EN regression), in which of those mental states the person was in at the time the data were obtained. We asked 40 participants to undergo fMRI while they decided whether to carry a hypothetical suitcase, which could have contraband in it, through a checkpoint. We varied the probability that the suitcase they carried had contraband, so that participants could be in a knowing situation (they knew the suitcase they were carrying had contraband) or a reckless situation (they were not sure whether there was contraband in it, but were aware of a risk of varying magnitude). We found that we were able to predict with high accuracy whether a person was in a knowing or reckless state, and this was associated with unique functional brain patterns. Interestingly, this high predictive ability strongly depended on the amount of information participants had available at the time the information about the risks was presented.

Materials and Methods

Experimental Details.

Participants. Forty participants were recruited according to a protocol approved by the Virginia Tech Institutional Review Board. Written informed consent was obtained from all participants. From these, one-half of the participants (n = 20; 10 females) were placed in the Contraband-First condition (see *Experimental paradigm* for details), whereas the other half (n = 20; 10 females) were placed in the Search-First condition. The mean age (± SD) for each group was 26.9 ± 10.2 and 32.9 ± 11.9 y old, respectively.

Experimental paradigm. Participants were told a cover story about carrying "valuable content" (such as documents, microchip processors, etc.), here referred to as "contraband," through a checkpoint (Fig. S1). Note that, although the instructions did not use the term contraband so as not to discourage participants that were averse to illegal behavior, we use the term here for convenience. In each trial, they were shown between one and five suitcases, only one of which actually contained contraband, and were asked whether they were willing to carry a suitcase randomly chosen from the group (Fig. S1*A*, *Left*). Hence, the number of suitcases shown represented the risk of carrying the target suitcase with contraband (Contraband Risk): if only one suitcase was presented, then the participants knew with certainty that the suitcase had contraband in it (knowing situation, P_{contr} = 1), whereas if more than one suitcase was presented, they were not sure whether the suitcase they were assigned contained contraband, but were aware of the risk (reckless situation, with P_{contr} = 0.5, 0.33, 0.25, or 0.2 of having contraband in the suitcase). Participants also had different probabilities of being caught (Search Risk), with the probability of being searched at the checkpoint ranging from P_{search} = 0 to 0.8 (symbolized by 10 tunnels, in which a proportion of them could be occupied by a "guard"; Fig. S1*A*, *Right*). One-half of the participants (n = 20) saw the probability of carrying a suitcase with contraband after already being shown the search risk (Search-First group), whereas the other half started by seeing the suitcases before being shown the search risk (Contraband-First group). See *Supporting Information* for details.

Data Analysis. See *Supporting Information* for details on the behavioral and fMRI data analyses. To perform the classification, we used an EN regression. The goal of this analysis was to understand whether, given a particular brain activation state, we could correctly predict which mental state the participant was in at the time the brain data were collected. Namely, we wanted to know whether we could disentangle whether the participant was in a knowing or a reckless situation. To achieve that, we used as a classifier the EN regression (see Fig. S2 and *Supporting Information* for step-by-step details). To assess the "significance" of the results, correcting for finite sample sizes (23), we ran a permutation test (*Supporting Information*).

Results

Behavioral Results. Behavioral data are presented in Fig. 1. Tests of within-subject effects from a mixed-model ANOVA revealed main effects for both Contraband Risk [$F_{(4,152)} = 20.7, P < 0.001$] and Search Risk [$F_{(4,152)} = 131.8, P < 0.001$] on the decision to carry the suitcase. Regardless of condition (Contraband-First or Search-First), as the likelihood of a suitcase containing contraband increased, decisions to carry the suitcase decreased. Similarly, regardless of condition, as the likelihood of being searched increased, decisions to carry the suitcase decreased. Furthermore, there was a significant Search Risk vs. Contraband Risk interaction [$F_{(16,608)} = 10.2$, $P < 0.001$]. A significant interaction was also observed between Search Risk and Condition [$F_{(4,152)} = 3.27, P = 0.013$] but not Contraband Risk and Condition [$F_{(4,152)} = 1.23$, $P = 0.302$], and a significant Contraband Risk by Search Risk by Condition interaction was observed [$F_{(16,608)} = 3.39$, $P = 0.002$]. Analysis revealed that the magnitude of the main effect of Search Risk was contingent on the order in which risk information was received. When collapsing across Contraband Risk, data show that, for identical degrees of Search Risk (00, 20, 40, 60, or 80%), seeing the search

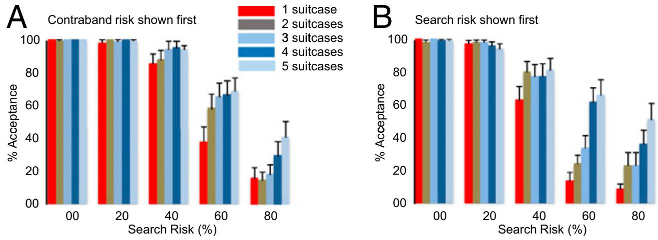

Fig. 1. Behavior summary. (*A*) Behavior for n = 20 participants seeing the contraband risk first (Contraband-First condition). The percentage of times the participant decided to carry the suitcase is on the y axis, whereas the Search Risk (proportion of tunnels occupied by a guard) is on the x axis. Colors code the Contraband Risk (number of suitcases presented, e.g., one suitcase: P_{contr} = 1; two suitcases: P_{contr} = 0.5, etc.). (*B*) Behavior for n = 20 participants seeing the search risk first (Search-First condition). Note the presence of a Search Risk by Contraband Risk interaction in both conditions, but stronger in the Search-First condition. Error bars represent SEM. See Table S1 for results of logistic regression.

risk before contraband risk resulted in fewer decisions to carry (mean ± SD = 99 ± 0.01%, 97 ± 0.02%, 76 ± 0.07%, 40 ± 0.23%, and 29 ± 0.16%), compared with seeing the Search Risk after the Contraband Risk (100 ± 0.00%, 99 ± 0.01%, 91 ± 0.04%, 59 ± 1.2%, and 24 ± 1.1%, respectively). This shows that, although the content and the level of risk associated with a single decision was identical, the order in which the information was received significantly altered choice behavior. Specifically, seeing the search risk before seeing the contraband suitcases typically decreased the choice to carry contraband suitcases. Similar results were obtained using a logistic regression (*Supporting Information*). As the order in which information was presented significantly affected behavior, these two groups/conditions will be analyzed separately. Finally, note that fewer decisions to carry contraband are made when individuals are in knowing as opposed to increasingly reckless situations [observe one-suitcase [red] trials relative to two-, three-, four-, and five-suitcase trials], indicating that the participant is indeed aware that he/she is carrying contraband.

Classifier Performance. Using the brain-imaging data from the participants in the Search-First condition (and only the trials in which participants decided to carry the suitcase), we were able to predict, with relatively high accuracy, whether the brain-imaging data corresponded to a knowing (Contraband Risk: $P_{contr} = 1$) or a reckless ($P_{contr} = 0.2$) situation (Fig. 2). The EN classifier had an out-of-sample average area under the curve (AUC) value of 0.789 (AUC values close to 1 indicate "perfect" classification, and close to 0.5 suggest random classification) and an average correct classification rate (CCR) of 71% (Fig. 2*A*). These values are significantly above chance, with *P* values obtained through a permutation test equal to $P_{perm} = 0.005$ (i.e., only 1 in 200 models run with shuffled labels had an AUC or CCR value as high or higher than these; see *Materials and Methods* and *Supporting Information* for details). This high accuracy was maintained even at the single-subject level and when using a more stringent, double–cross-validation procedure (see *Supporting Information* for details). We find several areas in the brain predictive of being in a knowing situation, namely dorsomedial prefrontal cortex (dmPFC) and medial orbitofrontal cortex (mOFC), middle and anterior cingulate cortex (ACC), bilateral superior temporal gyrus/temporoparietal junction (TPJ) and bilateral anterior insula (Fig. 2*B* and Table S2). Areas more predictive of being in a reckless situation were mainly in the occipital cortex (Fig. 2*C*). These brain areas were differentially activated in a knowing and reckless situation, and, together, the brain activity in them allowed predicting (significantly above chance) in which situation the person was.

If we do the same analysis using brain imaging data from the participants in the Contraband-First condition (i.e., at the time the contraband risk was being shown they had not seen the search risk yet), the results change. The accuracy of the EN classifier in distinguishing between the knowing and reckless condition drops to an out-of-sample average AUC value of 0.287 ($P_{perm} = 1$; Fig. 3*A*) and an average correct classification rate of 32.1% ($P_{perm} = 1$). For the knowing situation, the (right) TPJ also appears, and for the reckless situation identical occipital areas appear (Fig. 3*B* and Table S2). Note, however, that the coefficients associated with these voxels/areas have relatively small survival rates, indicating that, for many of the model runs, none of these voxels was very predictive of being in one state or another. Although the visual information presented in both conditions is identical, the lower predicting capability of the EN classifier in these data compared with the Search-First condition indicates that it is not the visual information in itself that drives the higher predictability of the model, and also that having or lacking complete information about both the contraband risk and the probability of getting caught (search risk) changes some of the brain patterns (or at least the strength of the signal) associated with it.

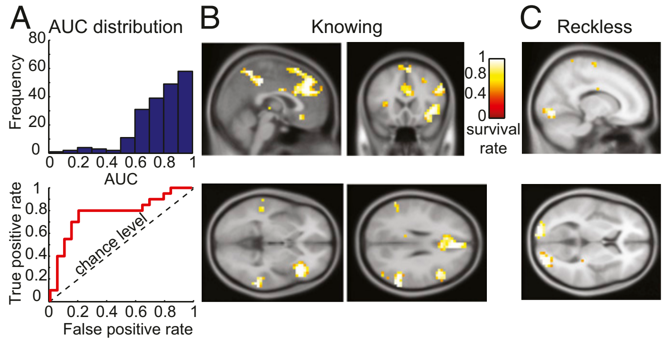

Fig. 2. The K/R distinction, for the Search-First condition. These results were obtained based on the brain state at the time that the contraband risk is revealed (suitcases shown), when the contraband risk is presented after the search risk (Search-First condition, $n = 20$). (*A*, *Top*) Distribution of cross-validated areas under the curve (AUCs). AUC values close to 1 indicate "perfect" classification, whereas those close to 0.5 suggest random classification. Forty iterations of a fivefold cross-validated EN regression were performed, resulting in the 200 AUC calculations plotted in the histogram (mean out-of-sample AUC = 0.79). (*Bottom*) Example of one receiver-operating characteristic (ROC) curve obtained, from which an AUC is drawn. The dashed line represents a "curve" from a model that would perform at chance level (hence the area under this "curve" is 50%, i.e., the AUC would be 0.5). ROC curves consistently above this dashed line are associated with AUC values higher than 0.5. (*B*) Areas predictive of being in a knowing situation ($P_{contr} = 1$). Represented is the (signed) survival rate for the voxels. The "signed survival rate" for a voxel is the proportion of times this voxel was used in the EN classifier (i.e., got coefficient values different from zero), multiplied by the sign of the average beta value for this voxel (see *Supporting Information* for details). Hence, absolute survival rate values closer to 1 mean that the voxel "survives" most of the cross-validated runs of the EN algorithm, indicating that this voxel is relevant in distinguish a knowing ($P_{contr} = 1$) from a reckless ($P_{contr} = 0.2$) situation. Voxels with a negative signed survival rate are shown, indicating regions predictive of being in the knowing situation (the base group in our model). (*C*) Areas predictive of being in a reckless situation ($P_{contr} = 0.2$; voxels with a positive survival rate). Each voxel's (signed) survival rate is overlaid on a sagittal (*B*, *Top Left*, $x = 2$; *C*, *Top*, $x = 14$), coronal (*B*, *Top Right*, $y = 20$), or axial (*B*, *Bottom*, $z = -2$ *Left*, $z = 26$ *Right*; *C*, $z = 6$) section of a 152-participant average T1 SPM brain template (minimum survival rate for the cluster's peak voxel of 0.5). The xjView program was used to display all of the brain figures.

The results obtained until now used $P_{contr} = 0.2$ (five suitcases presented) as the recklessness category. To analyze what happens to the EN model's classification accuracy when different contraband risks are used, we performed the same analysis but comparing the knowing situation (Contraband Risk: $P_{contr} = 1$) with different forms of recklessness, varying with the Contraband Risk ($P_{contr} = 0.5$, 0.33, 0.25, or 0.2; Fig. 4). We find that, for the Search-First condition, the EN classifier comparing the knowing with most other recklessness states also allowed for a significantly better than chance separation ability: for the EN classifier distinguishing one vs. three suitcases ($P_{contr} = 1$ vs. $P_{contr} = 0.33$), the average AUC was 0.924 and the CCR was 79.6% ($P_{perm} = 0$); and for the EN separating one vs. four suitcases ($P_{contr} = 1$ vs. $P_{contr} = 0.25$), the average AUC was 0.82 and the CCR was 75.7% ($P_{perm} = 0$). The performance of the EN classifier contrasting knowing with the recklessness state more near the knowing situation ($P_{contr} = 1$ vs. $P_{contr} = 0.5$) was slightly worse, with an average AUC value of 0.678 and a CCR of 55.3% ($P_{perm} = 0.13$ and $P_{perm} = 0.11$, respectively). On the other hand, for the Contraband-First condition, the EN classifier does not perform better than chance in distinguishing knowing from any of the recklessness situations: for $P_{contr} = 0.5$, the average AUC was 0.259 and the CCR was 32.2% ($P_{perm} = 1$ for both); for $P_{contr} = 0.33$ (one vs. three suitcases), the average AUC was 0.38 and the CCR was 35.3% ($P_{perm} = 0.96$ and $P_{perm} = 0.85$, respectively); and

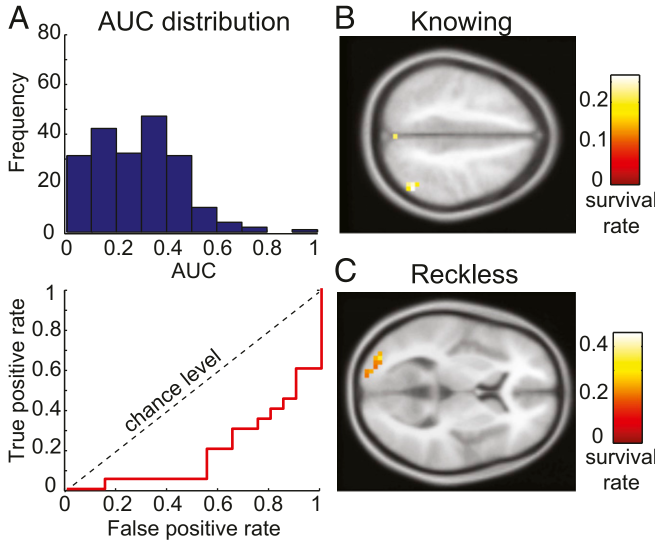

Fig. 3. The K/R distinction, for the Contraband-First condition. EN results obtained in a similar way to what was presented in Fig. 2, but when the contraband risk is presented before the search risk (Contraband-First condition, $n = 20$). (*A*, *Top*) Distribution of cross-validated AUCs averaged over 200 values. Mean AUC is 0.29. (*Bottom*) Example of one ROC curve. (*B*) Results obtained with the EN model. Here, the survival rates are low. (*Top*) Voxels more associated with a knowing situation are shown (negative surviving rate voxels). (*Bottom*) Voxels more associated with a reckless situation are presented (positive surviving rate). Each voxel's (signed) survival rate is overlaid on an axial section (*Top*, $z = 46$; *Bottom*, $z = 6$) of a 152-participant average T1 SPM brain template ($n = 20$; minimum survival rate for the cluster's peak voxel of 0.2). Note that, here, we had to reduce the survival rate plotting threshold, as no voxels appear with surviving rates higher than 0.5.

for $P_{contr} = 0.25$, the average AUC was 0.349 and the CCR was 34.2% ($P_{perm} = 0.98$ and $P_{perm} = 0.92$, respectively). The capability of the EN model to distinguish between a knowing and a reckless situation thus depended on both the degree of probability (Contraband Risk) and also the amount of information available to the participant (in terms of Search Risk) at the time that the knowing or reckless situation was being presented.

See *Supporting Information* for measures of single-subject precision, double–cross-validation, and several control analyses (Figs. S3–S7).

Discussion

In this paper, we set out to discover whether knowing and reckless mental states, as defined by the MPC, correspond to detectable different states in the human brain. Moreover, we wanted to know whether we could predict which of those mental states the person was in based only on the corresponding brain-imaging data. Using EN regression on brain-imaging data of people exposed to knowing or reckless scenarios, we found that knowing and reckless are indeed associated with distinct brain states. Moreover, it was possible to predict, with relatively high accuracy, which mental state the person was in. This study is a first step in understanding how the legally defined concepts of "knowledge" and "recklessness" map onto different brain states and shows, as a proof of principle, that it is possible to predict which legally defined mental state a person is in based only on imaging data.

The fact that our EN model was able to distinguish between a knowing or reckless state with higher than chance levels (although far from perfect) indicates that, at least for some conditions, the knowing and reckless states may indeed be correlated with, and so possibly realized by distinct states of the human brain. This predictive ability was consistently found for various recklessness states. However, for the recklessness state more near the knowing condition ($P_{contr} = 0.5$), the accuracy, even though relatively high, was not significantly better than chance. This suggests that the knowing/reckless (K/R) boundary may be more of a continuum, and that when recklessness involves awareness of probability values closer to those involved in the knowing situation, the K/R boundary may be at least difficult to distinguish, and perhaps even blurred. Also, the capacity of our model to distinguish the K/R states strongly depended on participants already having information about the risk of being searched. Together, these results are consistent with the idea that the human brain has a K/R boundary, but exactly how it is drawn may depend both on the distance between the knowing and reckless situations and on the amount of information available to the person (in terms of search risk) at the time the K/R situation is happening.

Observing the brain areas that were repeatedly used by the model to predict which situation the participant was in sheds light on what brain areas are differentially associated with a knowing and a reckless "mental state." One of the areas that appeared more predictive of being in a knowing situation was the anterior insula. This is in line with previous experiments implicating the anterior insula in risk and uncertainty representation (9, 10, 24). The insula was still differentially active even after taking into account potential effects related with "risk" in terms of uncertainty in reward, which is in accordance with studies suggesting that it may have a general role in uncertainty that is independent of the effects of reward (8, 10, 24). Another area more involved in knowing than in reckless states was the dorsomedial prefrontal cortex. The prefrontal cortex is generally associated with executive decisions and making computations (10, 25), including also the assessment of probabilities and uncertainty (10, 12). Interestingly, in our experiment this area seemed to be more engaged when the participants already had seen the search risk associated with the trial. This suggests that participants may be waiting to have all of the information available to them to compute their decision. Finally, we also obtained bilateral TPJ, which is known to be associated with moral decisions (26). Areas specifically more predictive of being in a reckless situation include the occipital cortex. Although this may be related to simple visual effects specific to our task, as more suitcases were presented on the screen in recklessness scenarios, these areas have also been associated with higher uncertainty in current information (likelihood), which is higher

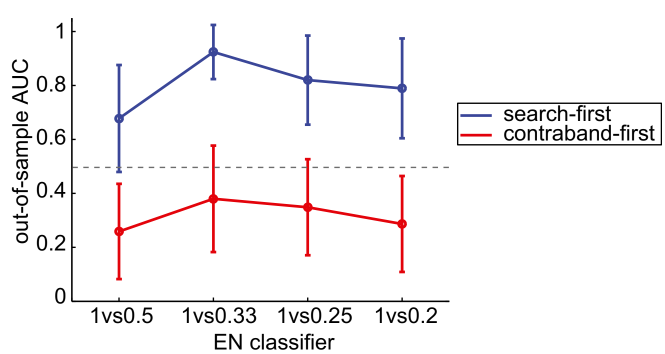

Fig. 4. Changing the boundary: The effect of different degrees of recklessness on the performance of the EN classifier. The out-of-sample performance of the EN classifier is shown, when the knowing situation is contrasted to different degrees of probability of carrying the target suitcase, that is, different forms of recklessness. The performance measure used here is the AUC. Represented are the mean AUC values ± SD (over the 200 runs) for each fitted EN classifier. The results for both conditions are represented, namely, the Search-First group (in blue) and the Contraband-First group (in red). The dashed gray line signals the expected mean AUC for a model that would perform at chance level (AUC = 0.5).

in a reckless scenario (10). Future studies may tell if the areas we found predictive of being in a knowing or reckless scenario generalize to other scenarios, for example, knowingly or recklessly evading taxes.

Although the EN model was able to classify with high accuracy a knowing or a reckless state in the Search-First condition, in the Contraband-First condition the model did not perform better than chance (even though the visual information was identical). This effect of order of presentation of information was also seen in behavior: seeing the search risk before contraband risk resulted in fewer decisions to carry. It is well known that human decision-making can be influenced by the manner in which options are presented (27, 28). Our results suggest that this is true not only for decisions involving multiple options but also for differing presentations of information related to a single decision. Alternatively, it may be that participants are waiting to have all of the information available to them to compute the associated contraband and search risks.

The following question can then be raised: Do the brain areas we are seeing correspond exclusively to knowing vs. reckless, or are they just representing the search risk (or their interaction)? In the inputs given to the model to distinguish K/R, search risk had already been averaged out [modeled in different betas on the same general linear model (GLM)]. Furthermore, if we analyze only the trials in which no search risk was present the same brain areas appear, indicating that they are differentially active in knowing vs. reckless even when no search risk exists (*Supporting Information*). Finally, extracting out the effects associated with the probability of being caught (i.e., searched while carrying contraband) still leads to the same results. Nevertheless, already having the information about search risk or not affects both the behavioral and the imaging results, hence search risk does matter in some way. In the real world, the probability of getting caught affects people's decision to commit, or not, a crime. It is then quite possible that the awareness of one's risk of being caught affects the manifestation of the culpable brain states themselves. Future studies could aim at understanding more precisely the effect of presentation of information and of search risk in the knowing and reckless brain states.

A word of caution: even though increased activations in the anterior insula, PFC, and TPJ were associated with a knowing scenario, this does not mean that this particular brain pattern/mental state could not appear in other situations, totally unrelated to the K/R distinction. For example, it may well be that they appear when assessing the probability of one event even if that event has no legal relevance. What it does mean is that, if the subject was either in a state of knowledge or reckless, then having this particular brain state increased the chances that the participant was in a state of knowledge (in contrast to recklessness).

To what extent is the difference between knowledge and recklessness, as defined by the law, the same as the difference between certainty and uncertainty? People are considered to act knowingly, under the law, when they are certain that their conduct is accompanied by a specific circumstance (in our experiment, that the suitcase contained contraband). In contrast, they are considered to act recklessly if they are aware of a "substantial and unjustifiable" risk that their conduct is accompanied by that circumstance, but unsure of it. So, the distinction between knowledge and recklessness is closely related to the ordinary distinction between certainty and uncertainty.

However, knowledge and recklessness are both likely to have more elements than certainty and uncertainty, respectively, have. There will be cases of certainty that are not cases of knowledge in the legal sense, and cases of uncertainty that are not cases of recklessness. The knowing and reckless mental states generally include an interpersonal relation, and they often include a moral dimension. The brain areas we found support this notion. Specifically, although the anterior insula has traditionally been implicated in uncertainty representation (among other things), albeit in nonlegally relevant settings (9, 10, 24), the TPJ has been more generally associated with moral decisions (26, 29). Note also that these areas appear even after abstracting out effects associated with Variance in Reward. Future studies could deploy a similar experimental setup and range of probabilities and choices but in a gambling scenario, or a scenario involving taking a ball from two urns, to see if similar areas are activated. Our prediction would be that, although the uncertainty-specific areas might be maintained (insula), others would not (e.g., TPJ).

The participants in this experiment, although more diverse than typical college student subjects in such experiments (30), are still not representative of the US population, let alone of the general human population. Limitations on generalizing the results obtained by this classifier include the fact that we have a small sample size ($n = 40$) and that the participant pool is restricted to the Roanoke/Blacksburg (Virginia) area. Furthermore, our experiment was done in a laboratory setting (with no real risk of going to jail), and participants were given the exact probabilities of events, whereas this may not be the case in "real life." Nevertheless, these results show a proof of concept: the knowledge and reckless mental states do seem to have distinct neural correlates, at least for some people and in a situation like the one portrayed in our experiment, and these neural correlates can be used to infer which state the person was in. More studies, from different independent laboratories, and with a broader participant pool are needed to analyze the generalizability of these findings.

Future studies could also look at the other MPC mental states not analyzed here, namely Purposeful and Negligent. Although we have shown here that a recklessness mental state could be distinguished from a knowing mental state, to confirm that recklessness is a mental state on its own future studies should see whether recklessness can be distinguished from Negligence using brain data alone. Similarly, future studies could look at the brain distinction between knowing and purposeful. The fact that typical jurors seem to be able to make these distinction behaviorally (1, 2) suggests this would be possible.

We conclude with some remarks about the potential legal relevance of our findings, recognizing that under no circumstances should legal practice be altered in the face of any single study, or even a small number of supporting studies. We want to first emphasize the negative; there are various tempting conclusions to reach that should be resisted. In particular, it would be absurd to suggest, in light of our results, that the task of assessing the mental state of a defendant could or should, even in principle, be reduced to the classification of brain data. For one thing, our capacity to classify participants' mental states depended on the collection of brain data at the time of a potentially criminal act. Obviously, in most cases, when someone is committing a crime they are not doing so while inside a scanner. We do not know whether it is possible, even in principle, to classify a person's mental state at a time that precedes the collection of brain data by minutes, hours, days, or even years, as is necessary in criminal trials. As it stands, our classifier represents a proof of concept, and not yet a usable tool. Future studies might assess whether this mental state can be recreated, for example by showing pictures of the circumstances of the potential crime, and whether a recreation of this kind would elicit particular brain states.

For another thing, our classifier's ability to predict the mental-state category of our participants was entirely dependent on our ability to classify the mental states of the participants in the "training" dataset without appeal to brain data. That is, our ability to classify on the basis of brain data was parasitic on our ability to conclude that, for instance, a participant who chose to carry the suitcase when only one suitcase was offered to him knew that he was carrying contraband. That conclusion was not reached through a study of his brain activations but, instead, through the commonsense interpretation of human behavior so familiar from everyday life. In addition, there are good reasons to believe that the legitimacy of our verdicts in criminal cases depends crucially on the fact, and the appearance, that the jury is making an unmediated judgement about the culpability of the defendant, rather than deferring to the results dictated by any nonhuman tool. That would be lost were anyone but the jury asked to assess the defendant's mental state.

However, this is not to suggest that our results have no legal significance. Legal scholars have argued about whether legally relevant mental states, such as those defined in the MPC, are arbitrary constructs or have some underlying resonance with actual psychological states. If the mental state categories are arbitrary constructs, then we should worry that differential punishments driven by differential mental-state classifications are equally arbitrary. Additionally, this is a source of potential worry, for arbitrarily constructed categories are at risk for interfering with the task of drawing merited distinctions; they sometimes, instead, may reflect biases or can even be used to serve the ends of the powerful. Our results suggest that the legally significant conceptions of knowledge (certainty that a particular circumstance exists) and recklessness (awareness of a possibility or probability that it exists) are distinctly represented in the human brain, and generalize existing results from the decision-making and neuroeconomics literature into the legal domain. These findings could therefore be the first steps toward demonstrating that legally defined (and morally significant) mental states may reflect actual, detectable, psychological states grounded in particular neural activities. Whether a reckless drug courier should be punished any less than a knowing one will of course always remain a normative question. However, that question may be informed by comfort that our legally relevant mental-state categories have a psychological foundation.

Also, even if several future studies confirm what we have observed here, that knowledge and recklessness are associated with different brain states, if human jurors cannot distinguish them behaviorally, then one may still ask whether they should be considered relevant to assessments of criminal liability. Our results here do not settle this question. However, they are suggestive. There could be no justice in punishing the knowing more harshly than the reckless, if there is, in fact, no difference in the minds of those whom we classify in one way and those we classify in another. However, our results suggest that there is indeed such a difference, and so it could be that we should work to help jurors to see the distinction, and classify defendants accurately under it, rather than abandoning it.

This work could also ultimately contribute to solving a more practical, but just as daunting, problem: We know almost nothing about the ways in which certain recognized mental disorders might impact the processing of information and the occurrence of the particular mental states that are inculpatory under the MPC. Currently, the law in many jurisdictions handles this problem by allowing defendants to introduce evidence of an alleged mental disorder (intoxication being the usual exception), and then letting the judge or jury speculate about whether that condition had any impact on the defendant's mental functioning at the time of the offense. So, for example, a defendant charged with a knowing crime might introduce evidence that he has a schizoaffective disorder and argue that that condition prevented him from acting knowingly or recklessly, despite the fact that we currently have little understanding about whether and under what conditions people suffering from schizoaffective disorder are able to process information about risks. Conversely, intoxication is generally not a defense to "recklessness" crimes, but many states allow evidence that a defendant was intoxicated at the time of an offense to show that he or she did not have the "knowledge" required for a "knowing" crime. Understanding more about the way our brains distinguish between legally relevant circumstances in the world has the potential to improve what, up until now, has been the law's guesswork about the ways in which certain mental conditions might impact criminal responsibility.

ACKNOWLEDGMENTS. We thank Frank Tong for useful discussions and all of the members from the Human Neuroimaging Lab, especially Alec Solway, Andreas Hula, and Sébastien Hétu, for helpful comments and discussion. We are also thankful for the support of the Wellcome Trust, the Kane Foundation, the Brown Foundation, and the National Institute on Drug Abuse. This study was supported by a grant from the John D. and Catherine T. MacArthur Foundation to Vanderbilt University, with a subcontract to Virginia Tech. Its contents do not necessarily represent official views of either the John D. and Catherine T. MacArthur Foundation or the MacArthur Foundation Research Network on Law and Neuroscience (www.lawneuro.org).

1. Shen FX, Hoffman MB, Jones OD, Greene JD, Marois R (2011) Sorting guilty minds. *NYU Law Rev* 86(5):1306–1360.
2. Severance LJ, Goodman J, Loftus EF (1992) Inferring the criminal mind: Toward a bridge between legal doctrine and psychological understanding. *J Crim Justice* 20(2):107–120.
3. Haynes JD, Rees G (2006) Decoding mental states from brain activity in humans. *Nat Rev Neurosci* 7(7):523–534.
4. Hughes V (2010) Science in court: Head case. *Nature* 464(7287):340–342.
5. Treadway MT, et al. (2014) Corticolimbic gating of emotion-driven punishment. *Nat Neurosci* 17(9):1270–1275.
6. Tobler PN, O'Doherty JP, Dolan RJ, Schultz W (2007) Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *J Neurophysiol* 97(2):1621–1632.
7. Huettel SA, Song AW, McCarthy G (2005) Decisions under uncertainty: Probabilistic context influences activation of prefrontal and parietal cortices. *J Neurosci* 25(13):3304–3311.
8. Preuschoff K, Bossaerts P, Quartz SR (2006) Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 51(3):381–390.
9. Preuschoff K, Quartz SR, Bossaerts P (2008) Human insula activation reflects risk prediction errors as well as risk. *J Neurosci* 28(11):2745–2752.
10. Vilares I, Howard JD, Fernandes HL, Gottfried JA, Kording KP (2012) Differential representations of prior and likelihood uncertainty in the human brain. *Curr Biol* 22(18):1641–1648.
11. Vilares I, Kording K (2011) Bayesian models: The structure of the world, uncertainty, behavior, and the brain. *Ann N Y Acad Sci* 1224:22–39.
12. d'Acremont M, Schultz W, Bossaerts P (2013) The human brain encodes event frequencies while forming subjective beliefs. *J Neurosci* 33(26):10887–10897.
13. McCoy AN, Platt ML (2005) Risk-sensitive neurons in macaque posterior cingulate cortex. *Nat Neurosci* 8(9):1220–1227.
14. Ahn WY, et al. (2014) Nonpolitical images evoke neural predictors of political ideology. *Curr Biol* 24(22):2693–2699.
15. Ryali S, Supekar K, Abrams DA, Menon V (2010) Sparse logistic regression for whole-brain classification of fMRI data. *Neuroimage* 51(2):752–764.
16. Haxby JV, Connolly AC, Guntupalli JS (2014) Decoding neural representational spaces using multivariate pattern analysis. *Annu Rev Neurosci* 37:435–456.
17. Gabrieli JD, Ghosh SS, Whitfield-Gabrieli S (2015) Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* 85(1):11–26.
18. Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. *Nat Neurosci* 8(5):679–685.
19. Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10(9):424–430.
20. Tong F, Pratte MS (2012) Decoding patterns of human brain activity. *Annu Rev Psychol* 63:483–509.
21. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc B* 67:301–320.
22. Whelan R, Garavan H (2014) When optimism hurts: Inflated predictions in psychiatric neuroimaging. *Biol Psychiatry* 75(9):746–748.
23. Combrisson E, Jerbi K (2015) Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J Neurosci Methods* 250:126–136.
24. Singer T, Critchley HD, Preuschoff K (2009) A common role of insula in feelings, empathy and uncertainty. *Trends Cogn Sci* 13(8):334–340.
25. Miller EK (2000) The prefrontal cortex and cognitive control. *Nat Rev Neurosci* 1(1):59–65.
26. Young L, Camprodon JA, Hauser M, Pascual-Leone A, Saxe R (2010) Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proc Natl Acad Sci USA* 107(15):6753–6758.
27. Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481):453–458.
28. De Martino B, Kumaran D, Seymour B, Dolan RJ (2006) Frames, biases, and rational decision-making in the human brain. *Science* 313(5787):684–687.
29. Moll J, Zahn R, de Oliveira-Souza R, Krueger F, Grafman J (2005) Opinion: The neural basis of human moral cognition. *Nat Rev Neurosci* 6(10):799–809.
30. Henrich J, Heine SJ, Norenzayan A (2010) The weirdest people in the world? *Behav Brain Sci* 33(2-3):61–83, discussion 83–135.
31. Deichmann R, Gottfried JA, Hutton C, Turner R (2003) Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage* 19(2 Pt 1):430–441.