# 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

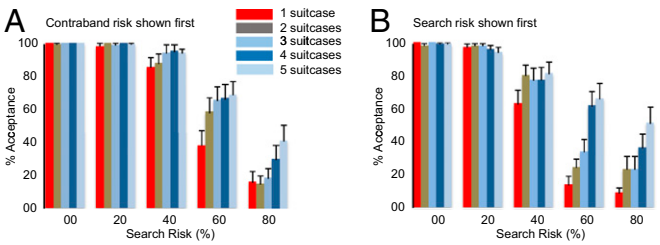


**Fig. 1.** Behavior summary. (*A*) Behavior for $n = 20$ participants seeing the contraband risk first (Contraband-First condition). The percentage of times the participant decided to carry the suitcase is on the *y* axis, whereas the Search Risk (proportion of tunnels occupied by a guard) is on the *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.

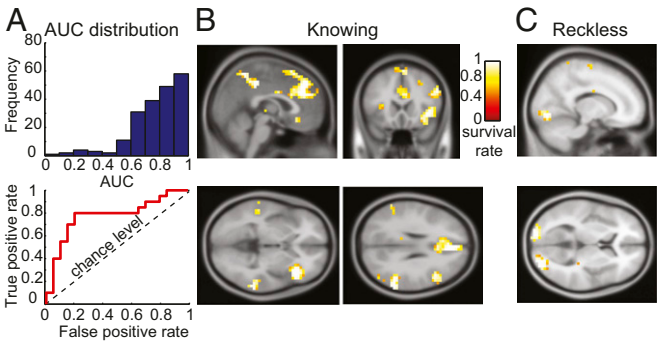


**Fig. 2.** The K/R distinction, for the Search-First condition. These results were obtained based on the brain state at the time that the contraband risk is revealed (suitcases shown), when the contraband risk is presented after the search risk (Search-First condition, $n = 20$). (*A*, *Top*) Distribution 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

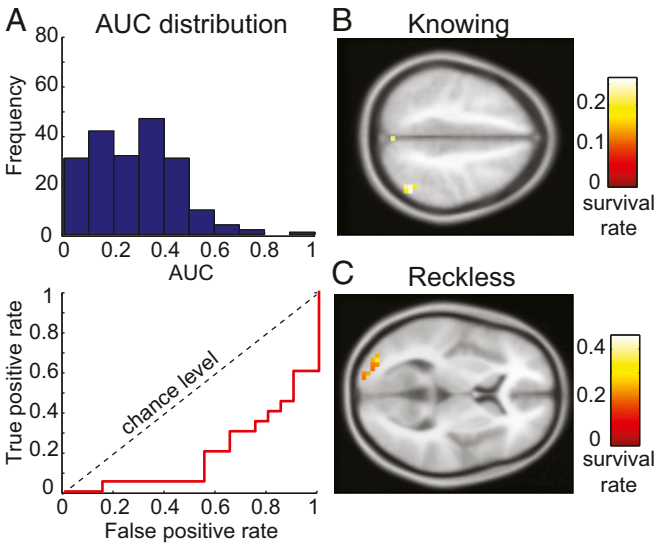


**Fig. 3.** The K/R distinction, for the Contraband-First condition. EN results obtained in a similar way to what was presented in Fig. 2, but when the contraband risk is presented before the search risk (Contraband-First condition, $n = 20$). (*A*, *Top*) Distribution of cross-validated AUCs averaged 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

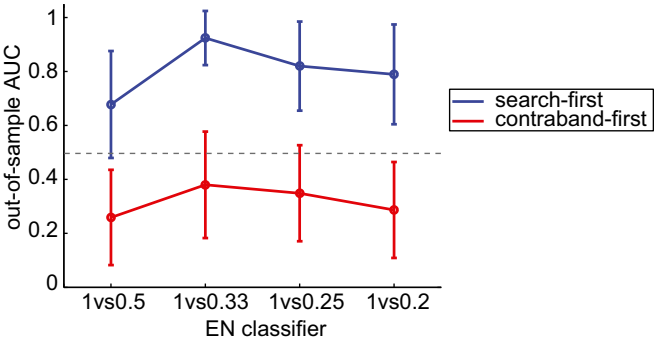


**Fig. 4.** Changing the boundary: The effect of different degrees of recklessness on the performance of the EN classifier. The out-of-sample performance of the EN classifier is shown, when the knowing situation is contrasted to different degrees of probability of carrying the target suitcase, that is, 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 non-legally 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.

# Supporting Information

## Vilares et al. 10.1073/pnas.1619385114

### SI Materials and Methods

#### Experimental Details.

***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, $P_{\text{contr}} = 1$), 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 $P_{\text{contr}} = 0.5, 0.33, 0.25$, 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 $P_{\text{search}} = 0$ 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).

The motivation to have a Contraband-First condition and a Search-First condition was to control for the order of presentation of the information. The motivation to not only change the contraband risk but to change the search risk as well was twofold. First, it enabled us to provide a more realistic setup—one that involved (as in real life) differing potential search risks. Second, it enabled us to vary search risks together with contraband risks, in a way that allowed us to disambiguate the effects of Contraband Risk, Expected Value, and Variance in Reward. Without variation in search risk, these would be perfectly correlated, and therefore indistinguishable.

Task payouts are shown in Fig. S1*B*. At the beginning of each trial a participant was endowed with \$6,000. After seeing the two types of risk information (Contraband Risk and Search Risk), participants decided whether or not to carry a suitcase. To carry a suitcase, participants had to pay \$500. If they decided not to carry, they had to pay \$1,500, leaving them with a trial total of \$4,500. If no decision was made (participants did not respond during the allocated time), participants lost \$2,500, leaving them with a trial total of \$3,500. These three costs associated with each trial (the cost to carry, the cost to reject, and the cost of inaction) were considered a "life tax" and were included (informed by pilot studies) to simulate motivations to act expressed by individuals engaging in the real-world analog of our experiment. If participants chose to carry a suitcase, if it contained contraband ("target suitcase") and if they were not searched, they received \$2,500 extra, getting a trial total of \$8,000 (the maximum possible payoff). However, if participants were searched while carrying the target suitcase containing contraband, they lost \$3,500, leaving them with a trial total of \$2,000 (the minimum possible payoff). If they chose to carry a suitcase and it did not contain contraband ("dummy suitcase"), they received no extra money, leaving them with a trial total of \$5,500, regardless of whether or not they were searched. Participants were not shown the results of individual trials. Hence, because no feedback was provided at the end of each trial, all trials were independent from one another. At the end of each trial, a computer simulated the outcome for that trial. At the end of the experiment, the computer randomly picked the payout of one trial and participants earned 1% of their trial total in US dollars. Hence, each participant received \$20 to \$80 at the end of the experiment, the exact value depending on the trial randomly chosen by the computer and the choices and outcome of that trial.

***Display/stimulus.*** The sequence of each trial was as follows (Fig. S1*C*): First, the contraband risk (Contraband-First group) or the search risk (Search-First group) was shown on the screen, for 2 s. Afterward, there was a blank screen for about 3 s (duration jittered between 2.5 and 3.5 s), and the search risk or contraband risk (respectively) was displayed, also for 2 s. Then another blank screen was shown (duration: average 3 s, jittered between 2.5 and 3.5 s), and a screen appeared indicating that the program was selecting which suitcase would be carried (duration: 0.75 s), after which another screen was shown asking the subject to choose to carry that suitcase or not. Participants had free time to make their decision, and they expressed their decision by pressing a button (representing either "yes" or "no"). If no button was pressed in the following 5 s, it was recorded as if no decision was made. After the decision was inputed (or the 5 s passed), the choice made by the participant was highlighted on the screen, and the computer calculated the results of that trial/round (not shown to the participant). Finally, about 0.75 s after the choice was submitted, participants were shown a screen with a fixation cross, which lasted for about 3 s (between 2.5 and 3.5 s, jittered), and then a new trial started.

Participants performed a total of 125 trials in the experiment (5 types of Search Risk × 5 types of Contraband Risk × 5 repetitions of each trial type). The exact sequence of trials shown was chosen randomly and differed among participants. The sequence of each trial can be seen in Fig. S1*C*.

#### Behavioral Data Analyses.

***Statistical tests.*** To determine whether there were significant differences in the choice to carry contraband based on the risk of having contraband, the risk of being searched, and the order of risk information received, a $5 \times 5 \times 2$ mixed-model ANOVA was conducted with Contraband Risk and Search Risk serving as within-group factors containing five levels each and the order of information seen serving as the two-level between-group factor (Contraband-First or Search-First). This analysis was done using the statistical software SPSS.

***Logistic regression.*** The participants' behavior was analyzed using logistic regression as implemented by the function glm in R (Team 2014). The dependent variable was a participant's response: Accept or Reject carrying the suitcase. The dependent variables were Contraband Risk, Search Risk, and the interaction Contraband Risk*Search Risk. The condition (Contraband-First or Search-First) was also entered as a categorical variable, with Contraband-First decisions coded as 0, and Search-First decisions as 1. The terms in the model were thus as follows: a constant, a dummy variable for Condition, Contraband Risk, Contraband Risk*Condition, Search Risk, Search Risk*Condition, Contraband Risk*Search Risk, and Contraband Risk*Search Risk*Condition.

#### fMRI Analysis.

***Scans and preprocessing.*** Anatomical and functional images were acquired on a Siemens 3-T Trio scanner at Virginia Tech Carilion

Research Institute in Roanoke, Virginia. A high-resolution (1.0 × 1.0 × 1.0-mm voxels) T1-weighted anatomical image was acquired for each participant using a magnetization-prepared rapid-acquisition gradient echo sequence. Functional images were acquired using echo-planar imaging. Slices were acquired at an angle 30° to the anterior–posterior commissure (31). The repetition time was 2 s, the echo time was 25 ms, and the flip angle was 90°. Thirty-four (interleaved) slices were acquired, resulting in functional voxels of size 3.4 × 3.4 × 4.0 mm. Preprocessing was done in SPM8. Briefly, the functional scans were slice-timing corrected, aligned to a functional average scan for that section, and unwarped. The anatomical scan was coregistered to the mean of the functional images. The anatomical image was segmented and normalized to the SPM templates, and the results used to normalize the functional images, which were resliced to 4.0 × 4.0 × 4.0-mm voxels. Finally, the functional images were smoothed using an 8-mm full-width half-maximum Gaussian kernel.

***General linear model.*** For every participant, a general linear model (GLM) was fitted to the participant's fMRI data (first-level analysis), using SPM8. A standard rapid event-related fMRI approach was used, in which the onset of each event type was convolved with a canonical hemodynamic response function and then regressed against the measured fMRI signal. The specific events regressed depended on what was being studied:

*i*) Main model: Resulting betas used as input for the different knowing vs. reckless comparisons (results from Figs. 2–4). The events modeled were as follow: five event types represented the different contraband risks (when one suitcase, two suitcases, three suitcases, four suitcases, or five suitcases were shown, corresponding to probabilities of 1, 0.5, 0.33, 0.25, and 0.2 of having the target suitcase with contraband), with the onset times modeled at the time in which the suitcase(s) were first shown, including only the trials in which the participant decided to carry the suitcase. The times in which the participant decided to not carry the suitcase were modeled as a different event, parametrically modulated by contraband risk. Five event types represented the different search risks (with 0, 2, 4, 6, or 8 of the 10 tunnels having a guard in it, corresponding to probabilities of 0, 0.2, 0.4, 0.6, and 0.8 of being searched), with the onset times corresponding to the moment in which the screen with the tunnels appears. These were not separated by carry/not carry responses because, for some conditions (e.g., when search risk was very high), some participants never decided to carry the suitcase. Finally, we modeled the event "choice submitted," with onset times modeled at the time in which an answer was given by the participant (shown by a button press).

*ii*) Model used to predict participant's choice (shown below in *SI Results and Discussion*): The events modeled were the five different contraband risks described above, modeled at the time contraband was shown; the five different search risks, modeled at the time search risk was presented; and two event types representing the two potential answers given by the participant for that trial ("yes," carry the suitcase; or "no," do not carry or did not decide), with onset times modeled at the time in which an answer was given by the participant (shown by a button press), for a total of 12 events.

*iii*) Model used to predict knowledge vs. recklessness when no search risk was present (Fig. S3*A*): 25 events, corresponding to all of the possible Contraband*Search Risk combinations (5*5), were modeled at the time contraband risk was presented. The Contraband*Search Risk events were not further stripped away from negative responses, as not all of the combinations had enough trials to be able to make predictions. Hence, the two event types representing the two potential answers given by the participant for that trial were also modeled.

*iv*) Model used to predict knowledge vs. recklessness at the time search risk was presented (shown below in *SI Results and Discussion*): Similar to the main model but with the different contraband risks modeled at the time search risk was first presented (and not when contraband was presented). Because modeling search risk at the same time would make for betas not uniquely defined, we did not include search risk events in this particular model.

*v*) Model used to predict knowledge vs. recklessness at the time choice was submitted (shown below in *SI Results and Discussion*): Similar to the main model but with the different contraband risks modeled at the time choice was submitted.

*vi*) Model used to predict knowledge vs. recklessness extracting away the effects associated with Variance in Reward (Fig. S3*B*): Similar to the main model but adding Variance in Reward as a parametric modulator to the Contraband Risk events. Then, for the EN model, only the main events (not the effects associated with the parametric modulator) are used.

*vii*) Model used to predict knowledge vs. recklessness extracting away the effects associated with Expected Value (Fig. S3*C*): Similar to the main model but adding Expected Value as a parametric modulator to the contraband risk events. It was not possible to add Expected Value and Variance in Reward at the same time to the model because, for some subjects, this resulted in multicollinearity.

*viii*) Model to distinguish the specific effects of Contraband Risk (probability of carrying the suitcase with contraband) from those associated with Expected Value and Variance in Reward (Figs. S4 and S5): one event was suitcases shown (modeled at the time suitcases were first presented), with three parametric modulators: Contraband Risk, Expected Value, and Variance in Reward. Another event was search risk shown (modeled at the time the screen with the tunnels appears), parametrically modulated by Search Risk; and two other events: when the participant responds yes and when the participant responds no (modeled at the time the motoric response is given). The individual contrasts at this level are then taken to the second level to do simple population responses to these effects (analysis done in SPM).

*ix*) Model used to predict knowledge vs. recklessness extracting away the effects associated with probability of being caught, that is, of being searched while carrying the target suitcase (the suitcase that has contraband on it): Similar to the main model but adding probability of being searched while carrying the target suitcase as a parametric modulator to the contraband risk events (Fig. S6).

*x*) Model used to predict knowledge vs. recklessness extracting away the effects associated with probability of obtaining the highest reward, that is, of not being searched while carrying the target suitcase (the suitcase that has contraband on it): Similar to the main model but adding probability of not being searched while carrying the target suitcase as a parametric modulator to the contraband risk events (Fig. S7).

*xi*) Model that compares knowing with a mixed of reckless states: Similar to the main model, but in which the different types of reckless trials are not modeled independently, that is, there is a contrast for knowing (one suitcase shown), and then only one contrast for reckless, which in this model is only classified as "more than one suitcase shown" and includes both two, three, four, and five suitcases shown. The contrast associated when the participant decides to not carry the suitcase after is also not parametrically modulated, so that there is no specific mention to the degree of recklessness in this particular model.

Besides these regressors (one for each event type), in all models there was a constant term (encoding the average blood oxygen

level-dependent response for that experiment/participant) and also six nuisance regressors, which corresponded to participant-specific head-movement parameters. Contrast images, derived from a pairwise contrast between each event type and an implicit baseline (equivalent to the beta image produced for that event), were then calculated for each event and used as baseline data for the elastic-net (EN) regression.

**EN Regression.**

***Data used to estimate the model.*** For each participant, we extracted two contrast/beta images (obtained by the GLM procedure outlined above): one belonging to the event type knowing (when only one suitcase was shown, and so participants knew that they had the target suitcase with contraband, i.e., $P = 1$ of having the target), and one belonging to a reckless situation (e.g., when five suitcases were presented, i.e., probability of having the target suitcase with contraband is $P = 0.2$). The brain data corresponding to each contrast condition are then reshaped into a vector (1 × number of voxels) and entered as a row in the data matrix. After doing this for each participant, the final data matrix that will be used for EN regression has $2n$ rows, with $n$ representing the number of participants and 2 representing the number of different events we are trying to predict; and $p$ voxels, corresponding to the number of voxels in the participant's contrast image. In our task, the matrices had 40 rows (20 participants in each group × 2 events) and 65,280 columns (each one corresponding to one voxel in the brain). We then took out the voxels that did not fall inside the brain, leading to a final number of voxels/columns of ∼21,000. Thus, for the EN model, we had 40 observations (2 observations per participant) and 21,000 predicting variables (i.e., features). Each observation was associated with one particular label (knowing or reckless), which we then tried to predict.

We chose to model knowledge vs. recklessness at the time the contraband risk is revealed to have a "cleaner" knowledge vs. reckless brain state, and so maximize our chances of observing these two brain states, should they exist.

***The EN regression.*** The EN regression (21) is a form of regularized linear regression that tries to minimize as follows:

$$\min_{\beta_0,\beta} \frac{1}{N}\sum_{i=1}^{N} l(y_i, \beta_0 + \beta^T x_i) + \lambda\left[(1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1\right],$$

where $y_i$ is the vector we are trying to predict (in our case, composed by two labels, knowing and reckless), $x_i$ are the predictor variables (in our model, the voxels), $N$ corresponds to the number of observations, $l(\ldots)$ is the loss function associated with the type of data we are using (e.g., Gaussian, binomial); $\beta$ corresponds to the coefficients of the model that we are trying to estimate (being $\beta_0$ the intercept), $\lambda$ (lambda) is a regularization parameter that controls the strength of the regularization (for high values of $\lambda$, all of the coefficients will tend to zero; for $\lambda = 0$, the EN regression becomes an ordinary least-squares regression), and $\alpha$ (alpha) is the EN mixing parameter, which varies from 0 to 1 and indicates the relative quantities of L2 norm penalized regression (ridge regression, which corresponds to an $\alpha = 0$) and L1 norm penalized regression (lasso regularized regression, which corresponds to an $\alpha = 1$). Having some kind of regularization of the $\beta$ coefficients/predictors is very important to minimize overfitting in the case in which the number of predictors ($p$) is much greater than the number of participants ($n$) (i.e., $p >> n$) (14), as is the case with fMRI data (if we assume that each voxel is a variable). Ridge regression ($\alpha = 0$) does not make variable selection, so although it shrinks the $\beta$ coefficients, it keeps all of them, hence it becomes harder to disentangle which predictor variables are important. Lasso regression, on the other hand, does make variable selection but it does not allow for several correlated coefficients to remain in the final model, even if these coefficients are also important for the model. Furthermore, it never retains more variables than the number of observations. The EN penalized regression, by mixing these two types of penalization, is able to make variable selection while at the same time allowing for clustering of potential relevant variables.

***Parameter selection and model fitting.*** To fit the EN penalized logistic regression, we used the glmnet package for Matlab (web.stanford.edu/~hastie/glmnet_matlab/), which also included a function to do cross-validated model fitting (cvglmnet.m) and a function to make predictions (glmnetPredict.m). To select the two tuning parameters for the EN, the mixing parameter ($\alpha$) and the overall complexity parameter ($\lambda$), we first estimated $\alpha$ and, after $\alpha$ was fixed, $\lambda$ was selected (14) (Fig. S2). To estimate $\alpha$, similar to the procedure used by Ahn et al. (14), we conducted a grid search over different values of $\alpha$ (from 0 to 1, in small increments).

*i*) For each potential $\alpha$ value, we did the following procedure:

  *a*) Fivefold cross-validated EN fitting, using the cvglmnet.m function. What this function does, step by step, is as follows:

    *i*) At the particular value of $\alpha$ chosen, it first fits the EN model paths to get the $\lambda$ sequence.
    *ii*) It then divides the data randomly into five groups, or folds.
    *iii*) For each value of $\lambda$, it fits a model (in this case, a logistic model) using penalized maximum likelihood, that is, it fits an EN model, using only 80% of the data (four of the five folds previously made).
    *iv*) It then tests the model on the left-out fold, and computes the corresponding minimum binomial deviance. It does that for each $\lambda$.
    *v*) Steps *i*, *ii*, *iii*, and *iv* are then repeated five times (one for each fold), and the average binomial deviance over the five repetitions is computed for each $\lambda$.

  *b*) For each of these runs (performed by cvglmnet), we then saved the minimum average binomial distance obtained with the model.
  *c*) For each $\alpha$, we repeated 40 times steps *a* and *b*, and calculated the mean of the minimum average binomial deviance over the 40 repetitions.

*ii*) We then chose the $\alpha$ that minimized this value.
*iii*) After choosing $\alpha$, we proceeded to estimate $\lambda$, and the $\beta$ coefficients of the model. We repeated the following procedure 40 times:

  *a*) We ran a fivefold cross-validated EN regression (identical to step 1a), using the $\alpha$ obtained from steps *i* and *ii*.
  *b*) From it, we obtained the $\lambda$ value that minimizes the binomial deviance ($\lambda_{min}$).
  *c*) We also obtained the indices associated with the division of the data in five folds.
  *d*) Then, for $\lambda_{min}$, we:

    *i*) Refitted an EN regression using only four folds of the data (using the division made by the *cvglmnet* function) and extracted the corresponding $\beta$ coefficients for all regressors (voxels);
    *ii*) Recorded, for each regressor (voxel), if it "survived" in the current run, that is, if its $\beta$ value is not zero.
    *iii*) We then used the resulting EN model and tested it on the "left-out" fold, obtaining the "fitted probabilities" associated with each observation in the left-out fold, and also the corresponding most likely classification (i.e., if the observation most likely came from a knowing or reckless situation). From

these, we computed both the correct classification rate (CCR; how often did it accurately guess the situation the observation belonged to) and the area under the curve (AUC) of the receiver-operating characteristic curve. Note that the CCR and the AUC computed this way represent the out-of-sample, cross-validated performance of the model, as the observations in the left-out fold were not used to calculate the EN model, and are suggestive of how well the model would generalize.

*iv*) Repeated steps *i* through *iii* five times, each time leaving out a different fold.

*iv*) After doing steps *a* through *d* 40 times, we then calculated a matrix with the "signed survival rate" for each voxel. For that, we calculated the "survival rate" of each voxel, meaning how many times that voxel "survived" on the total of the 200 runs done (40 iterations × 5 folds), and we multiplied this survival rate with the sign of the mean $\beta$-coefficient values obtained for that voxel.

*v*) We then projected the signed survival rate of each voxel back into the brain. An average positive survival rate indicates voxels more predictive of being in the reckless state (i.e., the estimated coefficient values associated with those voxels were on average positive), whereas a negative survival rate indicates voxels more predictive of being in a knowing situation (which, in our model, was the "baseline" label). Higher survival rates indicate that the voxel was used frequently to distinguish the knowing and reckless situations.

*vi*) To calculate the overall performance of the model, we made the average of the 200 AUCs and CCRs obtained during step 3 (40 iterations × 5 folds per iteration).

***Leave-one-subject-out cross-validation.*** To obtain single-subject accuracy, we performed the same general steps as outlined above but using data from only $n - 1$ participants each time. Specifically, in each iteration, we leave the data of one participant out and do a fivefold CV to obtain both $\alpha$ and $\lambda$ (using the same general steps as outlined above). We then fit an EN model using the data of the $n - 1$ participants and the parameters ($\alpha$ and $\lambda$) obtained through fivefold CV, extract the corresponding betas, and test the data on the left-out participant. This procedure is then repeated $n$ times (equal to the number of participants).

***Double–cross-validation procedure.*** To obtain a very stringent measure of out-of-sample performance, we did a "double–cross-validated" procedure. Namely, we divided the participants in half, randomly, and built an EN classifier using only one-half of the data (doing fivefold cross-validation of the parameters within that data, following the same general steps outlined above), and then tested the resulting model on the untouched other half of the data. This other half of the data then serves as a completely independent dataset, as it was not used at all to train the EN classifier. This process is then repeated for the other half and the results averaged. For the outer fivefold inner-fivefold double cross-validation, the procedure is exactly the same, just that the participant's data are divided in five (instead of two) groups, and the EN classifier fitted on four-fifths of the data (using fivefold CV within it to fit it) and tested on the remaining one.

***Permutation test.*** If there is no real information present in the data, a binary classifier (as in the example here, given that we are trying to classify two different labels, knowing and reckless) should give, on average, a correct classification rate of 50% and an AUC of 0.5. Hence, anything above that should indicate that the model is performing better than chance. However, these values only hold for infinite sample sizes, and for small sample sizes the values can be higher (or lower) by chance (23). To assess the "significance" of the results we obtained in the models, we ran a permutation test, in which, for each permutation run, we followed the exact same procedure as outlined above, but in which we shuffled the labels corresponding to each observation (i.e., in our experiment, the labels that said if the data belonged to a knowing or a reckless situation). We iterated this procedure 200 times. Then, to obtain the "*P* value" for a particular AUC/CCR value, we calculated the proportion of iterations that had an AUC/CCR that high or higher. The results obtained with the CCR were identical to the ones presented in Combrisson and Jerbi (23) and very close to the theoretical values (based on a binomial cumulative distribution function, for number of observations = 40 and 2 different groups to be distinguished), with "statistical significance" (i.e., a *P* value of 0.05) achieved on average around AUC = 0.70/CCR = 60%. Note: for the permutation test, we fixed $\alpha = 0.5$ (due to time constrains). Redoing all of the analysis reported in the text using $\alpha = 0.5$ leads to the same basic results.

## SI Results and Discussion

**Behavioral Analysis—Logistic Regression.** The logistic regression over participants' behavior revealed that, even when the monetary gain/loss potential was equal, individuals who processed the likelihood of being searched by guards first (Search-First condition), followed by the likelihood that a suitcase contained contraband, were less likely to carry a contraband suitcase compared with those who processed the same information in the opposite order (Contraband-First condition; Table S1).

More precisely, logit regression analysis showed (*i*) a significant main effect of Search Risk; (*ii*) a significant Contraband Risk by Search Risk interaction; and (*iii*) a significant Condition by Contraband Risk by Search Risk interaction: The effect of Contraband Risk became greater as Search Risk increased, and this effect was bigger when the Search Risk was seen first. This order-dependent behavioral effect can be seen as a temporally extended framing effect. 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. In the context of the current task, it is plausible that the likelihood of being searched is a more aversive signal compared with the likelihood that a case being carried might contain contraband. Our results indicate that this signal, when processed before further information arrives, increases the impact of Contraband Risk, making knowing (that is, Contraband Risk = 1) even more salient.

**Control Analyses.** To further confirm that the high performance of the EN classifier in the Search-First condition is not just driven by differences in visual information, we reran the obtained EN classifier (distinguishing one vs. five suitcases) but taking away all of the surviving voxels that were part of the occipital/visual cortex. The resulting EN classifier maintained its high predicting ability, having an out-of-sample average AUC value of 0.834 ($P_{\text{perm}} = 0.005$) and an average CCR of 70.6% ($P_{\text{perm}} = 0.005$), suggesting once more that it is not the visual information driving this high performance of the classifier.

**Single-Subject Precision.** To obtain a measure of single-subject precision, we fitted an EN classifier using a leave-one-subject-out procedure on top of the fivefold cross-validation (see *SI Materials and Methods, EN Regression, Leave-One-Subject-Out Cross-Validation* for details). We found that, for the Search-First condition, the EN classifier was able to predict with high accuracy whether the brain data corresponded to a knowing (Contraband Risk: $P_{\text{contr}} = 1$) or a reckless ($P_{\text{contr}} = 0.2$) situation. The EN classifier had an out-of-sample mean AUC of 0.944 ($P_{\text{perm}} = 0$) and a mean correct classification rate of 81.8% ($P_{\text{perm}} = 0$). For the Contraband-First condition, the EN classifier had an out-of-sample mean AUC of 0.499 ($P_{\text{perm}} = 0.33$) and a mean correct classification rate of 50% ($P_{\text{perm}} = 0.1$). Thus, in our Search-First condition, we were able

to obtain a high single-subject precision in distinguishing a knowing from a reckless scenario.

**Half-Split/Double–Cross-validation.** Within each group, we also split the participants in half, randomly, and built an EN classifier using only one-half of the data (and doing fivefold cross-validation of the parameters within that data). We then tested the resulting model on the untouched other half of the data. Using this extra, more stringent, analysis, we still observe a higher-than-chance prediction accuracy in the Search-First condition. Specifically, the EN classifier achieved an out-of-sample mean AUC of 0.765 ($P_{perm} = 0$) and a mean correct classification rate of 73.9% ($P_{perm} = 0$). For the Contraband-First condition, the EN classifier had an out-of-sample mean AUC of 0.503 ($P_{perm} = 0.46$) and a mean correct classification rate of 50.5% ($P_{perm} = 0.15$). Similarly, when we split the data into five groups, train the classifier on four-fifths of the data (using fivefold CV within this group), and then test on the left-out data, we also obtain good prediction accuracies. This approach yielded a mean AUC of 0.803 ($P_{perm} = 0$) and a mean CCR = 73.3% ($P_{perm} = 0$). Hence, the higher-than-chance prediction accuracies observed in the Search-First condition when classifying knowing vs. reckless are maintained even when using very stringent analyses.

**Additional Analyses.** The lack of predictive power for the EN in the Contraband-First condition is somewhat surprising, given that the only thing that changed between conditions was the order of presentation of information. A trivial explanation could be that the functional imaging data of one or more participants' in the Contraband-First condition is corrupted. If that were the case, then it would not be possible to obtain any good predictive models with this dataset. However, when predicting participants' decision to carry the suitcase (see below), the EN model performed with very high accuracies (AUC > 0.9) both in the Search-First group but also in the Contraband-First group. This indicates that there is some other reason for the low performance in the Contraband-First condition. Our behavioral analysis revealed that, although the content and the level of risk associated with a single decision were identical, the order in which the information was received significantly altered choice behavior (see *Results*, *Behavioral Results*, and also below). Thus, in our task, the order in which participants received information about contraband and search risk affected both their behavior and the corresponding imaging data.

**Knowing vs. Recklessness (Degree of Recklessness Not Specified).** We were also interested in understanding whether knowing could be broadly distinguished from reckless, even if no additional information about the degree of recklessness were given. For that, we built an EN classifier in which we contrasted knowing states with a general reckless state (which included all reckless states), not specifying the degree of recklessness (see *SI Materials and Methods* for details). We found that, for the Search-First condition, we were still able to predict with high accuracy whether the brain data corresponded to a knowing or reckless scenario, obtaining an average AUC value of 0.872 and a CCR of 75.9% ($P_{perm} = 0.005$ and $P_{perm} = 0$, respectively). Thus, knowing seemed to be broadly distinguishable from reckless, even in the absence of information about the degree of recklessness.

**Knowing vs. Recklessness (No Search Risk).** To understand better whether the identified brain pattern was specifically associated with the extent of knowledge the participant had about the existence of contraband in the suitcase or whether the brain state only exists when the participant knows that there is a risk of getting searched, we reran the analysis but using only the knowledge and recklessness trials in which search risk was 0 (no guards on the tunnels). Hence there was no probability of getting caught while carrying the contraband. Behaviorally, we can also see that they were aware the search risk was 0, as participants almost always decided to carry the suitcase when there was no risk of being searched (Fig. 1). We find that, for the Search-First condition, the EN model using only the no-search-risk trials did a good job in classifying a knowledge vs. recklessness scenario, giving a mean AUC = 0.832 and a mean CCR = 70.9% ($P_{perm} = 0$ for both). Moreover, the brain areas obtained related to a knowing scenario were identical to the ones obtained in the full model [mPFC, cingulate cortex, insula, temporoparietal junction (TPJ); Fig. S3*A*]. Hence, the brain pattern we identified associated with the state of knowledge appears even if there is no threat.

**Knowing vs. Recklessness or Just Difference in Number of Suitcases?** Finally, we wanted to know if the high accuracy results obtained with the classifier on the Search-First condition were mainly due to it being able to distinguish any linear increase in the number of suitcases being presented, and not to anything specific about the knowing/reckless distinction. If that is the case, then the classifier should perform similarly in distinguishing, say, two vs. five suitcases or distinguishing one vs. four suitcases. However, the EN classifier built to disentangle between two suitcases vs. five suitcases being presented (using the same procedure as before) did not perform better than chance, having an out-of-sample average AUC value of 0.243 and an average CCR of 30.6% ($P_{perm} = 1$ for both). Contrast these values with the average values obtained at distinguishing between one vs. four suitcases (AUC = 0.82 and CCR = 75.7%, $P_{perm} = 0$ for both). This once more indicates that the obtained high accuracy results (in the Search-First condition) are not simply due to a visual increase in the number of suitcases being presented, and suggests that there may be something special about the knowing/reckless distinction.

**Brain Areas Specifically Associated with Contraband Risk, Expected Value, and Variance in Reward.** To try to understand whether the brain differential activations we observed between knowing and reckless were related to differences in Contraband Risk (knowing, $P = 1$ of existence of contraband in the suitcase vs. reckless, $0 < P < 1$, aware of a possibility but not certainty of the existence of contraband) or whether they were just related to differences in Expected Value or Variance in Reward ["risk" as defined by the neuroeconomics literature (7–9)], we reran the same analyses but extracting out the effects associated with either of these factors (by modeling them separately at the first-level GLM model and not including them in the input data for the EN regression). We find that we still have a higher-than-chance accuracy in predicting a knowing vs. reckless scenario, with the EN regression having an out-of-sample CCR = 71.4% and an AUC = 0.791 ($P_{perm} = 0.005$; Fig. S3*B*) when extracting out the potential effects associated with Variance in Reward, and an out-of-sample CCR = 71.7% and an AUC = 0.792 ($P_{perm} = 0.005$; Fig. S3*C*) when excluding the effects associated with expected reward. Furthermore, if we do a simple GLM modeling independently Contraband Risk, Variance in Reward, and Expected Value, we see that the brain pattern we observed to be related to the knowing/reckless distinction continues to be specifically associated with Contraband Risk (probability of carrying the suitcase with contraband; see also Figs. S4 and S5). The EN model also performs well after taking out potential effects associated with the probability of being "caught," that is, of being searched while carrying the suitcase with contraband (CCR = 71.8%, AUC = 0.792, $P_{perm} = 0.005$; Fig. S6), or the probability of getting the highest reward, that is, carrying the target suitcase and not being searched (CCR = 71.7%, AUC = 0.792, $P_{perm} = 0.005$; Fig. S7). Together, this indicates that the general brain pattern we see associated with the knowledge/recklessness distinction seem to be specifically related to the probability of carrying the suitcase that has contraband (Contraband Risk) and cannot be explained only

by differences in expected value, variance in reward, fear of being searched, or expectation of highest reward.

**Simple GLM Results—Search Risk and Contraband Risk.** To understand which brain areas are specifically involved in signaling search risk and contraband risk, we added separate regressors for Contraband Risk and Search Risk in a traditional GLM model and analyzed the areas that were parametrically correlated with them.

The areas we found positively correlated with Search Risk were mainly in the visual cortex, namely the calcarine sulcus [$P < 0.05$, family-wise error (FWE) corrected]. Areas more active with decreasing Search Risk include the bilateral TPJ, dorsolateral prefrontal cortex (DLPFC), and middle temporal gyrus ($P < 0.05$, cluster size FWE corrected). Interestingly, the bilateral TPJ and middle temporal gyrus were more active with decreasing Search Risk both in the Search-First condition but also in the Contraband-First condition, although somewhat less significant. In comparison, for the Contraband Risk (associated with the knowledge/recklessness distinction), whereas in the Search-First condition Contraband Risk was robustly positively correlated with a whole range of areas, namely dorsomedial prefrontal cortex (dmPFC), bilateral insula, bilateral TPJ, middle temporal gyrus, DLPFC, and cingulate cortex ($P < 0.05$, FWE corrected), for the Contraband-First condition no areas appear, even at a very lenient threshold ($P < 0.01$, uncorrected). Thus, the order in which information was presented also had a strong effect on the brain activations associated with Contraband Risk.

Note that, although not identical, there was some overlap between the areas negatively correlated with Search Risk and the areas involved in distinguishing knowing vs. reckless (e.g., bilateral TPJ, DLPFC). These areas appear even though Contraband Risk and Search Risk were modeled as independent events within the same GLM, indicating that they are independently activated by both Contraband Risk and Search Risk. These areas could be generally involved in signaling risk. However, if this were the case, then they should be positively (and not negatively) correlated with Search Risk. Both the TPJ and the DLPFC have been associated with moral decision-making (26, 29). It may be that, as the Search Risk decreases and it becomes more easy to carry contraband across borders, the choice is less of a risky one (how likely am I to get caught?) and more a moral one (should I do it?). Alternatively, it may also well be that these areas are specifically involved in signaling certainty, be it knowing that there was contraband in the suitcase or that they would not be searched. An interesting future study to tackle this issue would be to have participants do a task in which both the risks and potential rewards are similar to the ones adopted in this experiment, but in which there was no legal/contraband-carrying cover story; hence there would be no potential moral dilemma.

**Prediction Using Different Points in Time Within a Trial.** We chose to model knowledge vs. recklessness at the time the contraband risk is revealed to have a "cleaner" knowledge vs. reckless brain state, and so maximize our chances of observing these two brain states should they exist. To analyze whether the prediction capability of the EN would hold when other times are used, we fitted new models in which knowledge ($P_{contr} = 1$) and recklessness ($0 < P_{contr} < 1$) were modeled either at the time Search Risk was being presented or at the time choice was being submitted. For the model comparing knowledge ($P_{contr} = 1$) vs. recklessness ($P_{contr} = 0.33$) using the time in which Search Risk was presented, the EN model did not perform better than chance: for the Contraband-First condition, we obtained a mean AUC = 0.475 ($P_{perm} = 0.54$) and CCR = 37.2% ($P_{perm} = 0.67$); and the Search-First condition had a mean AUC = 0.384 ($P_{perm} = 0.99$) and a CCR = 35.7% ($P_{perm} = 0.85$). If we use the times in which choice was submitted, the model also does not perform better than chance: for the Contraband-First condition, there was a mean AUC = 0.448 ($P_{perm} = 0.62$) and CCR = 38.3% ($P_{perm} = 0.51$); and for the Search-First condition, mean AUC = 0.5 ($P_{perm} = 0.42$) and CCR = 42.8% ($P_{perm} = 0.34$). During those times, the brain may be more engaged in processing the current information (search risk or making a decision), and the information about knowledge and recklessness may not be as salient. Thus, the maximum predictability of the model was achieved when modeling the results at the time the contraband risk is presented.

**Predicting Choice.** We can use the EN model approach to try to predict participant's choice (decision to carry or not carry the suitcase) based on their brain data at the time the decision is submitted. We found that, for both conditions, the EN model was able to predict, with high accuracy, the decision of the participant. For the Search-First condition, the EN model had a mean AUC = 1 and a correct classification rate of 99.8% ($P_{perm} = 0$); and for the Contraband-First condition, the EN classifier had a mean AUC = 0.968 and a mean CCR = 90.9% ($P_{perm} = 0$). Doing a similar model but in which we try to predict the participant's choice based on brain data at the time the choice screen is first shown (i.e., before they press a button signaling their choice) also yields high performance results: the Search-First condition had a mean AUC = 1 and CCR = 100% ($P_{perm} = 0$), and the Contraband-First condition had a mean AUC = 1 and CCR = 99.4%. Although these very high performance accuracies are, at least in part, likely due to differential motoric activations (as participants had to press a button in the right hand to say yes and in the left hand to say no), these results serve as a good EN tool validation, showing that it is possible to use the EN regression in both conditions (Search-First and Contraband-First) to distinguish, with very high accuracy, brain states belonging to two different scenarios.

**Brain Areas Specifically Associated with Contraband Risk, Expected Value, and Variance in Reward.** To understand whether the brain pattern we observed used in distinguishing knowing vs. reckless was specifically associated with awareness of Contraband Risk (probability of carrying contraband), or whether it was just related to Expected Value or Variance in Reward, we performed a GLM modeling independently Contraband Risk, Expected Value, and Search Risk. We found that higher Contraband Risk (higher probability) remained positively associated with increased activations in the dmPFC, middle and anterior cingulate cortex, bilateral middle temporal gyrus, bilateral TPJ, and bilateral anterior insula; and negatively associated with bilateral activations in the occipital cortex ($P < 0.05$, FWE corrected; Fig. S4*A*). There were no areas surviving correction for multiple comparisons for higher Variance in Reward, and just the right TPJ was significantly correlated with lower Variance in Reward ($P < 0.05$, FWE corrected; Fig. S4*B*). For Expected Value, the "traditional" areas appeared, namely ventromedial prefrontal cortex (vmPFC) and ventral striatum ($P < 0.05$, FWE corrected; Fig. S4*C*). Both Contraband Risk and Expected Value seem to, independently, activate the superior temporal gyrus, TPJ, and part of the medial PFC (although mainly nonoverlapping areas; Fig. S5). Thus, although some brain areas were activated by several factors, the brain pattern chosen by the EN regression to distinguish knowing vs. reckless seems to be more specifically associated with Contraband Risk.

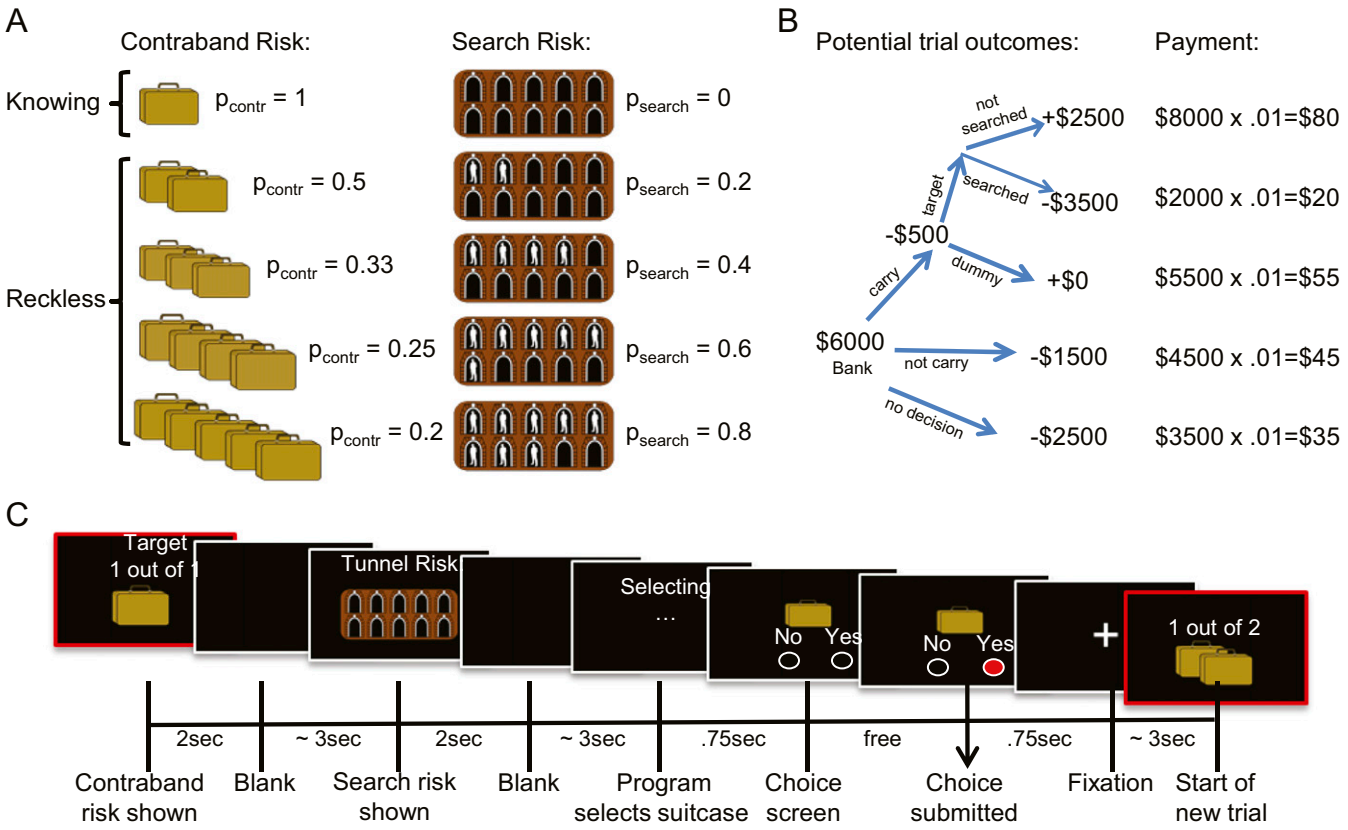


**Fig. S1.** Experimental design. (*A*) A display of the different scenarios participants were exposed to. (*Left*) Probability of carrying contraband: Participants were presented with one to five suitcases and asked whether they wanted to carry the suitcase. Only one of the suitcases was the "target" suitcase, supposedly containing contraband, and the other ones were dummies. Hence, when only one suitcase was shown, there was certainty that it was the target suitcase containing contraband (i.e., $P_{contr} = 1$). This corresponds to the knowing situation. As the number of suitcases shown increases, there is a lower probability that the person will carry the target suitcase ($P_{contr} = 0.5$, 0.33, 0.25, or 0.2 of having contraband in the suitcase). All of these other situations (with probability of carrying the target suitcase lower than 1) correspond to a reckless situation. (*Right*) Different potential search risk levels. This risk represents the probability of being searched by a "guard." If the participant is searched and has the target suitcase, he or she incurs a big penalty. The proportion of tunnels with guards indicates the search risk level. (*B*) Schematic display of the potential decisions and corresponding outcomes that can occur in a given trial. (*C*) Sequence of events shown to the participants in a typical trial. One-half of the participants ($n = 20$) were shown the contraband risk first, and then the search risk (Contraband-First group), and the other half of the participants were shown the search risk first (Search-First group).

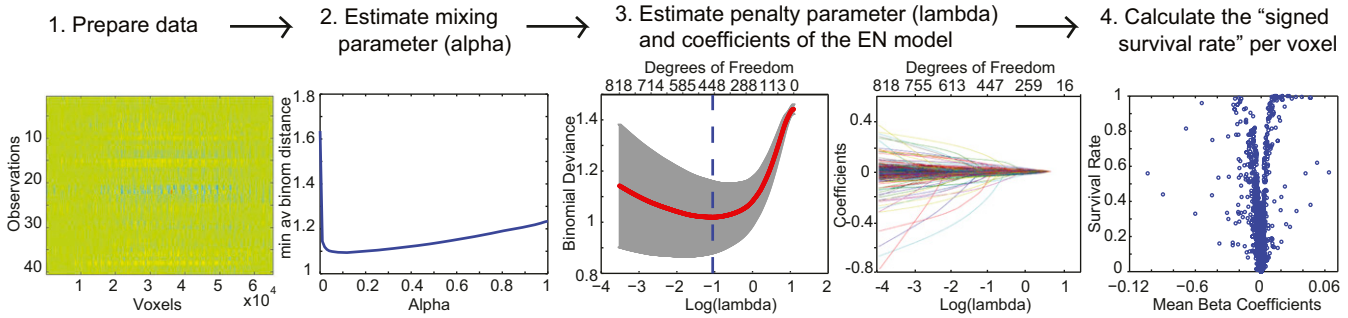


**Fig. S2.** Schematic representation of how the elastic-net (EN) model is implemented. Step 1: Preparing the matrix of data that is going to be used for the EN. For each participant, a simple general linear model (GLM) is first run, which includes as events the occasions in which they see "one suitcase" and also when they see "five suitcases." The corresponding beta maps are extracted and converted in rows, so that each participant is represented in two rows, one for "one suitcase" (knowing) and one for "five suitcases" (reckless). The length of each row corresponds to the total number of voxels present (e.g., ~65,000), minus the columns associated with NaN values (corresponding to nonbrain data). Step 2: Using the matrix obtained in 1, a grid search over $\alpha$ is performed, and the $\alpha$ that minimizes the average minimum binomial distance is chosen. Step 3: The EN is fitted over a range of $\lambda$ values, and for each $\lambda$ a series of coefficients is obtained (one coefficient per voxel). Then, the $\lambda$ that minimizes the binomial deviance is chosen (minimum $\lambda$). For this $\lambda$ ($\lambda_{min}$), we extract the corresponding coefficients and register which voxels "survived" (i.e., had coefficients different from zero), using fivefold cross-validation. This procedure is then repeated 40 times, for a total of 200 runs (fivefold cross-validation × 40 repetitions). Step 4: After all of the iterations are done, the survival rate is calculated for each voxel, multiplied by the sign of the average coefficient value over 200 CV runs, leading to a "signed survival rate" value per voxel. See *SI Materials and Methods* for details.

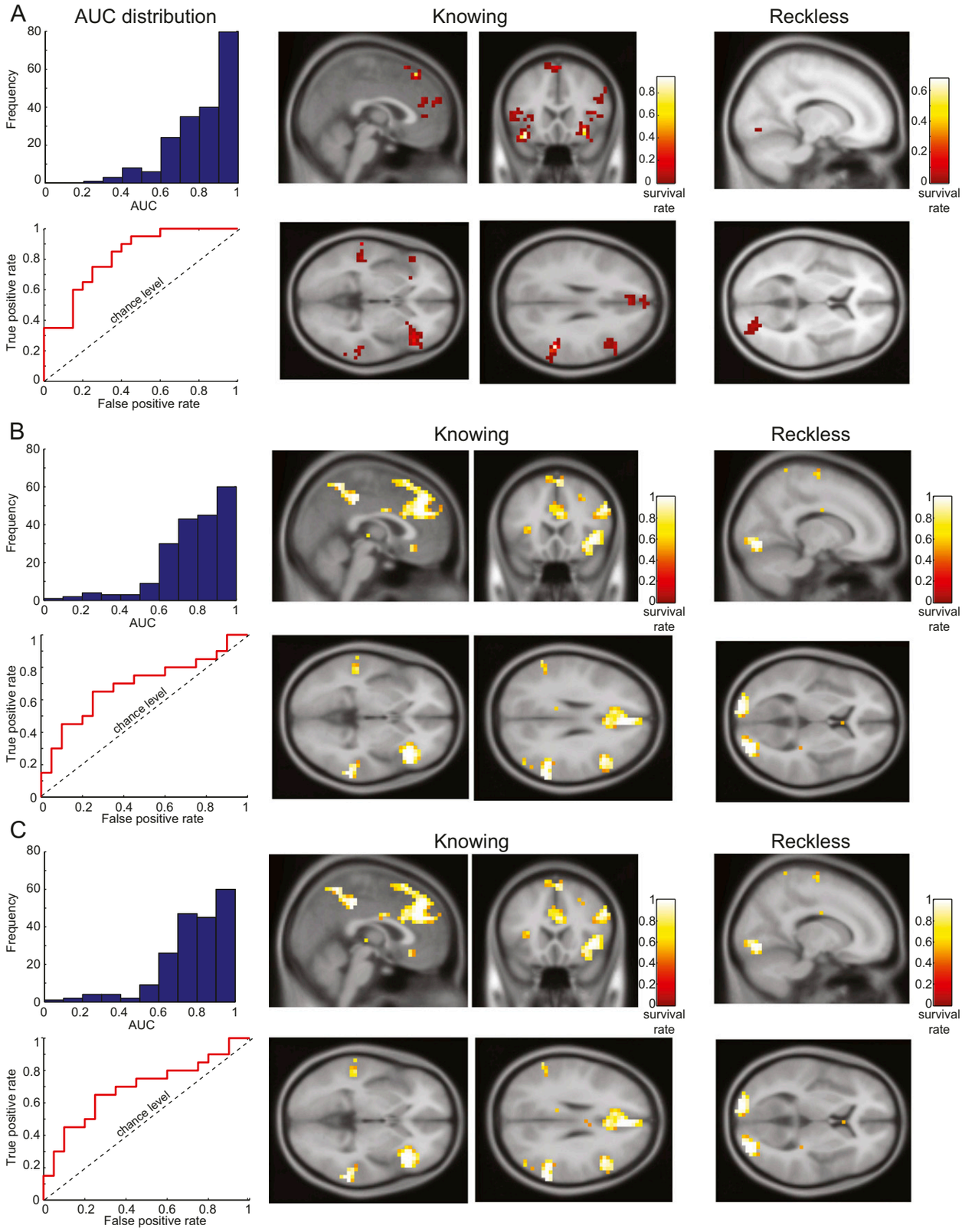


**Fig. S3.** Control figures. EN results obtained and displayed in a similar way to what was presented in Fig. 2, but using different baseline models. (*A*) Results obtained with the EN when only using the trials that have no search risk (no guards present in the tunnels). The model led to good prediction accuracies, with a CCR = 70.9% and an AUC = 0.832 ($P_{perm} = 0$). (*B*) EN results when using as a baseline the betas from a GLM in which the effects of Variance in Reward were modeled separately and not included in the EN model. The model still retained its high prediction ability (CCR = 71.4% and AUC = 0.791, $P_{perm} = 0.005$). (*C*) EN results when using as a baseline the betas from a GLM in which the effects of Expected Reward were modeled separately and not included in the EN model. The model revealed similar prediction accuracies (CCR = 71.7% and AUC = 0.792, $P_{perm} = 0.005$). In all figures: (*Top Left*) Distribution of cross-validated areas under the curve (AUCs) averaged over 200 values. (*Bottom Left*) Example of one ROC curve. (*Middle*) Voxels more associated with a knowing situation are shown (negative surviving rate voxels; $x = 2$, $y = 20$, $z = -2$ and $z = 26$). (*Right*) Voxels more associated with a reckless situation are presented (positive surviving-rate voxels, $x = 14$; $z = 6$). Each voxel's (signed) survival rate is overlaid on an axial section of a 152-participant average T1 SPM brain template higher ($n = 20$; with a minimum survival rate for the cluster's peak voxel of 0.5; for *B* and *C*, only the voxels with a minimum survival rate of 0.5 are shown).

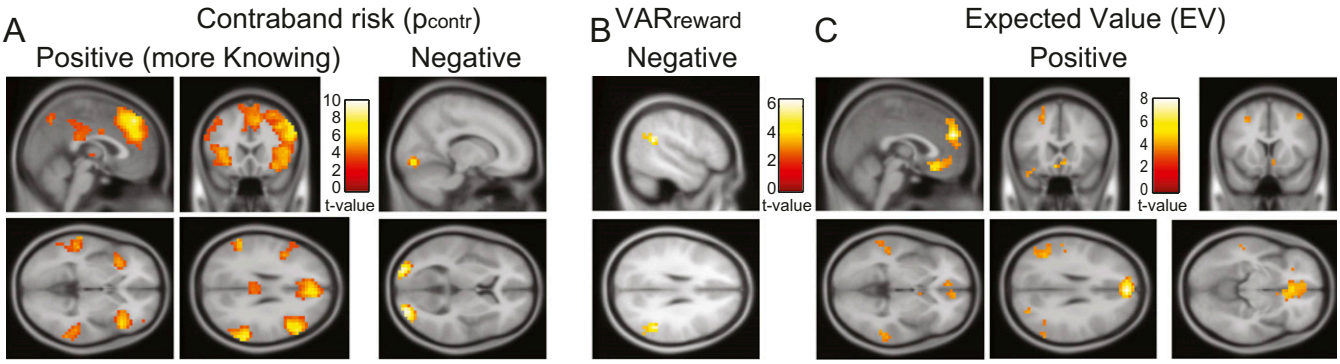


**Fig. S4.** Effects of Contraband Risk, Variance in Reward, and Expected Value (GLM results), for the Search-First condition ($n = 20$). (*A*) Represented are the areas positively (*Left*; $x = 2$, $y = 20$, $z = -2$ and $z = 26$) and negatively (*Right*; $x = 14$, $z = 6$) associated with the probability of carrying the suitcase that has contraband in it (Contraband Risk). (*B*) Represented are the areas negatively associated with Variance in Reward ($x = 50$, $z = 26$). No areas positively associated with Variance in Reward survived correction for multiple comparisons. (*C*) Represented are the areas positively parametrically associated with Expected Value (*Left*, $x = 2$ and $z = -2$; *Middle*, $y = 20$ and $z = 26$; *Right*, $y = 12$ and $z = -10$). No areas negatively associated with Expected Value survived correction for multiple comparisons. All areas represented survive correction for multiple comparisons ($P < 0.05$, FWE corrected either at the peak and/or cluster level, but areas displayed at $P < 0.001$, uncorrected; minimum cluster size, five voxels). Activations overlaid on a 152-participant average T1 SPM brain template.

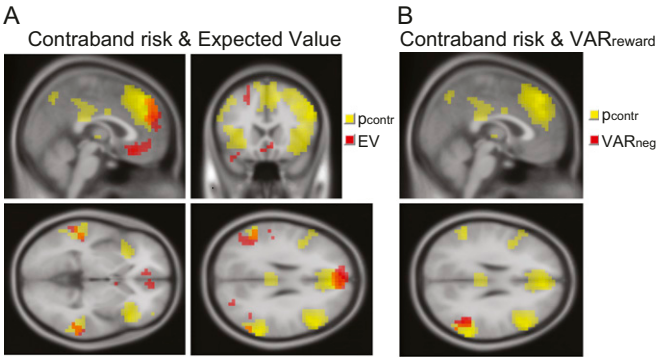


**Fig. S5.** Effects of Contraband Risk, Variance in Reward, and Expected Value (common regions), for the Search-First condition ($n = 20$). (*A*) The brain regions positively associated with Contraband Risk (in yellow) and with Expected Value (in red) are overlaid together. Common regions are shown in orange (original threshold for GLM maps of $P = 0.001$, uncorrected, $x = 2$, $y = 20$, $z = -2$ and $z = 26$, shown in Fig. S4). (*B*) Both the regions positively associated with Contraband Risk (in yellow) and negatively associated with Variance in Reward (in red) are overlaid. Common regions are shown in orange ($x = 2$ and $z = -2$, original threshold for GLM maps of $P = 0.001$, uncorrected). Activations overlaid on a152-participant average T1 SPM brain template.

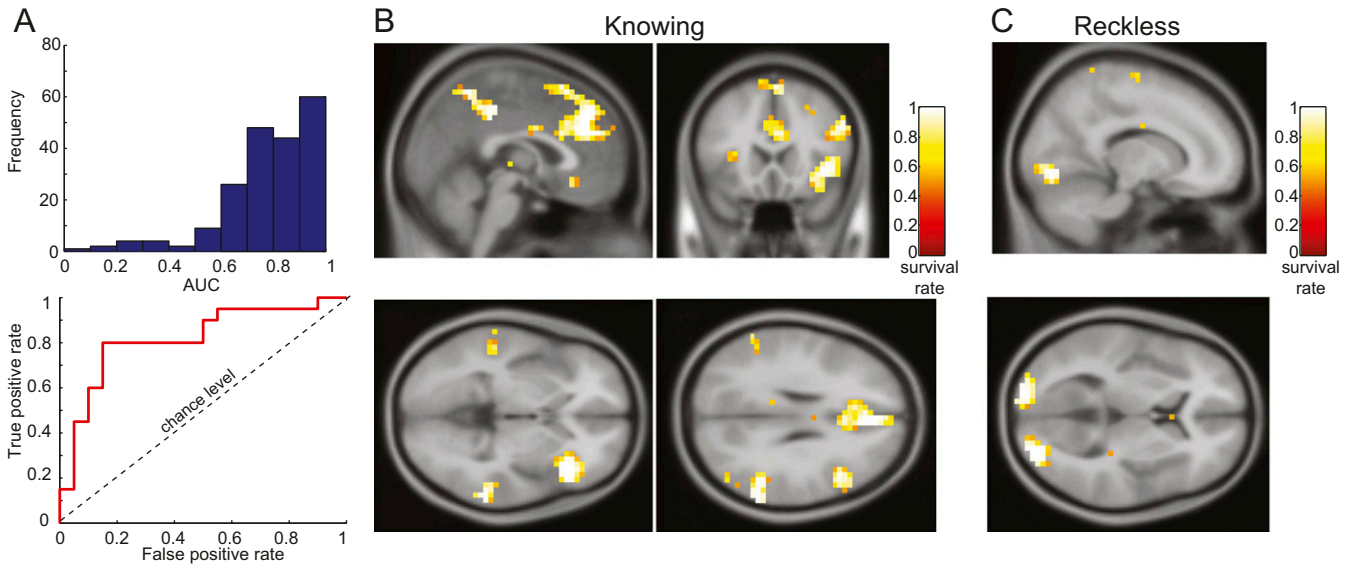


**Fig. S6.** The knowing/reckless distinction, extracting out parametric effects associated with the probability of being caught. EN results when using as a baseline the betas from a GLM in which the effects of probability of being caught (searched and with target suitcase) were modeled separately and not included in the EN model (Search-First condition, $n = 20$). The model led to higher-than-chance prediction accuracies, with a CCR = 71.8% and an AUC = 0.792 ($P_{perm} = 0.005$). (*A*, *Top*) Distribution of cross-validated areas under the curve (AUCs) averaged over 200 values. (*Bottom*) Example of one ROC curve. (*B*) Voxels more associated with a knowing situation are shown (negative surviving-rate voxels; $x = 2$, $y = 20$, $z = -2$ and $z = 26$). (*C*) Voxels more associated with a reckless situation are presented (positive surviving-rate voxels, $x = 14$; $z = 6$). Each voxel's (signed) survival rate is overlaid on an axial section of a 152-participant average T1 SPM brain template higher (with a minimum survival rate for the cluster's peak voxel of 0.5).

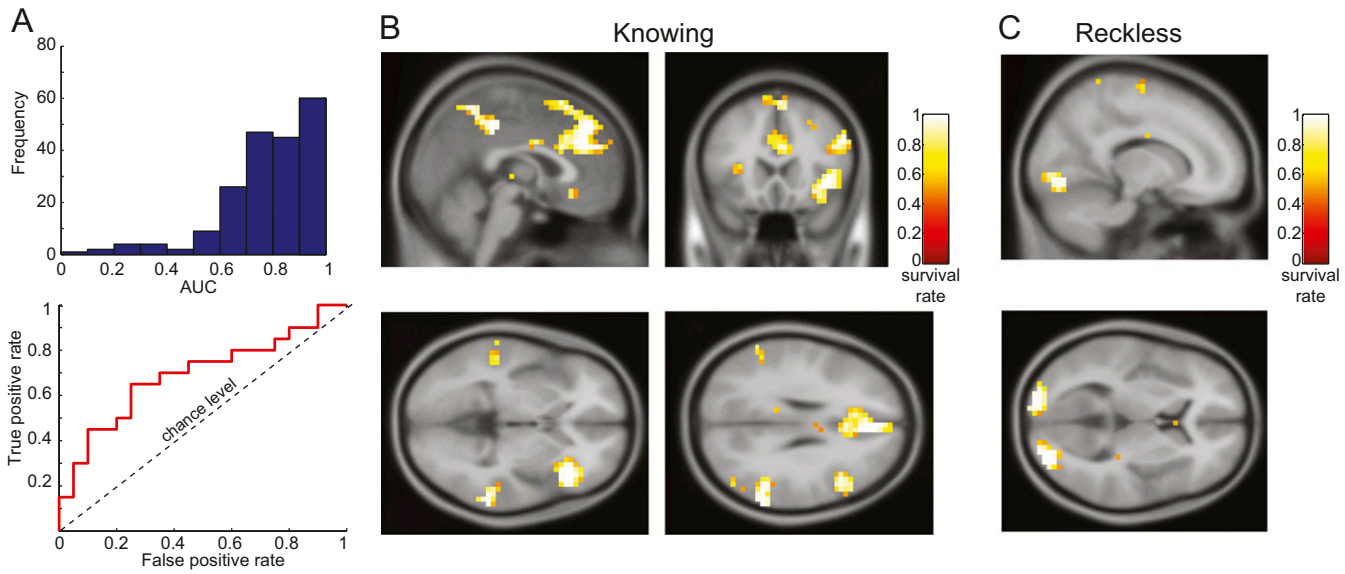


**Fig. S7.** The knowing/reckless distinction, extracting out parametric effects associated with the probability of obtaining the highest reward. EN results when using as a baseline the betas from a GLM in which the effects of probability of obtaining the highest reward (not searched and with target suitcase) were modeled separately and not included in the EN model (Search-First condition, $n = 20$). The model still retained its high prediction ability (CCR = 71.7% and AUC = 0.792, $P_{perm} = 0.005$). (*A*, *Top*) Distribution of cross-validated areas under the curve (AUCs) averaged over 200 values. (*Bottom*) Example of one ROC curve. (*B*) Voxels more associated with a knowing situation are shown (negative surviving-rate voxels; $x = 2$, $y = 20$, $z = -2$ and $z = 26$). (*C*) Voxels more associated with a reckless situation are presented (positive surviving-rate voxels, $x = 14$; $z = 6$). Each voxel's (signed) survival rate is overlaid on an axial section of a 152-participant average T1 SPM brain template higher (with a minimum survival rate for the cluster's peak voxel of 0.5).

**Table S1. Behavioral regression**

| Coefficient | Estimate | SE | *z* score | Pr(>\|*z*\|) |
|---|---|---|---|---|
| Intercept | −4.1977 | 0.2333 | −17.992 | <2e-16 |
| Search Risk | 5.3727 | 0.3985 | 13.482 | <2e-16 |
| Contraband Risk | −0.4321 | 0.4967 | −0.870 | 0.38436 |
| Search*Contraband | 2.8369 | 0.8799 | 3.224 | 0.00126 |
| Search*Condition | −0.1337 | 0.2640 | −0.507 | 0.61238 |
| Contraband*Condition | −0.1235 | 0.4776 | −0.258 | 0.79603 |
| Search*Contraband*Condition | 2.5885 | 1.0680 | 2.424 | 0.01536 |

Logistic regression of response (positive coefficient means for increased probability of rejecting carrying the suitcase) on Search Risk (0, 0.2, 0.4, 0.6, 0.8), Contraband Risk (0, 0.2, 0.25, 0.33, 0.5, 1), the interaction of Search and Contraband Risk (Search*Contraband), interaction of Search Risk and Condition (Condition 1 is contraband risk seen first, and Condition 2 is search risk seen first), interaction of Contraband Risk and Condition, and the three-way interaction of Contraband Risk, Search Risk, and Condition.

**Table S2. Surviving voxels**

| | | | Coordinates | | | |
|---|---|---|---|---|---|---|
| Conditions | H | Area | *x* | *y* | *z* | Survival rate |
| Search-First knowing | | | | | | |
| | L/R | Dorsomedial prefrontal cortex | 2 | 40 | 46 | 1 |
| | L/R | Anterior cingulate cortex | 6 | 36 | 26 | 1 |
| | R | Insula | 42 | 24 | 2 | 1 |
| | R | Temporoparietal junction | 58 | −48 | 30 | 1 |
| | L | Temporoparietal junction | −58 | −48 | 30 | 1 |
| | R | Middle temporal gyrus | 58 | −40 | 2 | 1 |
| | L/R | Middle cingulate cortex | 6 | −36 | 38 | 1 |
| | R | Dorsolateral prefrontal cortex | 46 | 16 | 26 | 0.95 |
| | L/R | Medial orbitofrontal cortex | 2 | 28 | −14 | 0.85 |
| | L | Middle temporal gyrus | −54 | −32 | −2 | 0.81 |
| | L | Insula | −30 | 20 | 10 | 0.67 |
| reckless | | | | | | |
| | R | Visual cortex | 26 | −88 | 6 | 1 |
| | L | Visual cortex | −18 | −96 | 2 | 0.99 |
| | R | Supplementary motor area | 14 | −12 | 70 | 0.73 |
| Contraband-First knowing | | | | | | |
| | R | Temporoparietal junction | 50 | −60 | 46 | 0.26 |
| reckless | | | | | | |
| | L | Visual cortex | −30 | −88 | 2 | 0.46 |
| | R | Orbitofrontal cortex | 38 | 32 | −14 | 0.30 |
| | R | Visual cortex | 34 | −84 | 14 | 0.28 |

Represented are the *x*, *y*, *z* coordinates for (one of the) peak voxels within each brain area used by the EN regression to predict knowing vs. reckless. H, Hemisphere. Coordinates are listed in standard Montreal Neurological Institute space.