

2016

Predicting the Knowledge–Recklessness Distinction in the Human Brain

Iris Vilares
University College London

Michael J. Wesley
University of Kentucky College of Medicine


Woo-Young Woo-Young Ahn
The Ohio State University

Richard J. Bonnie
University of Virginia - School of Law

Morris B. Hoffman
Second Judicial District Court Judge, State of Colorado

See next page for additional authors

Follow this and additional works at: http://scholarship.law.upenn.edu/faculty_scholarship

 Part of the [Criminal Law Commons](#), [Law and Psychology Commons](#), [Other Neuroscience and Neurobiology Commons](#), [Public Law and Legal Theory Commons](#), and the [Science and Technology Law Commons](#)

Recommended Citation

Vilares, Iris; Wesley, Michael J.; Woo-Young Ahn, Woo-Young; Bonnie, Richard J.; Hoffman, Morris B.; Jones, Owen D.; Morse, Stephen J.; Yaffe, Gideon; Lohrenz, Terry; and Montague, Read, "Predicting the Knowledge–Recklessness Distinction in the Human Brain" (2016). *Faculty Scholarship*. 1743.

http://scholarship.law.upenn.edu/faculty_scholarship/1743

Authors

Iris Vilares, Michael J. Wesley, Woo-Young Woo-Young Ahn, Richard J. Bonnie, Morris B. Hoffman, Owen D. Jones, Stephen J. Morse, Gideon Yaffe, Terry Lohrenz, and Read Montague

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. Morseⁱ, Gideon Yaffe^{j,2}, Terry Lohrenz^b, and P. Read Montague^{a,b,2}

^aWellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, United Kingdom; ^bVirginia Tech Carilion Research Institute, Virginia Tech, Roanoke, VA 24016; ^cDepartment of Behavioral Science, University of Kentucky College of Medicine, Lexington, KY 40506; ^dDepartment of Psychology, Ohio State University, Columbus, OH 43210; ^eInstitute of Law, Psychiatry and Public Policy, University of Virginia, Charlottesville, VA 22903; ^fSecond Judicial District (Denver), State of Colorado, Denver, CO 80202; ^gVanderbilt Law School, Vanderbilt University, Nashville, TN 37203; ^hDepartment of Biological Sciences, Vanderbilt University, Nashville, TN 37203; ⁱUniversity of Pennsylvania Law School, University of Pennsylvania, Philadelphia, PA 19104; and ^jYale 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.

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

²To 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 (\pm 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. S1A, 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. S1A, 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

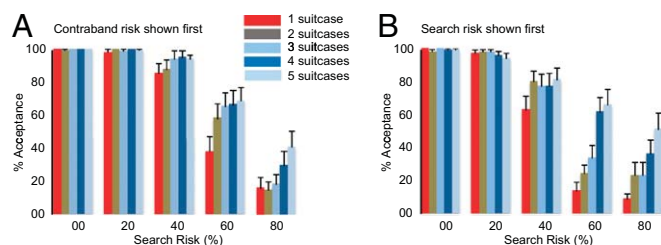


Fig. 1. Behavior summary. (A) Behavior for $n = 20$ participants seeing the contraband risk first (Contraband-First condition). The percentage of times the participant decided to carry the suitcase is on the y axis, whereas the Search Risk (proportion of tunnels occupied by a guard) is on the x axis. Colors code the Contraband Risk (number of suitcases presented, e.g., one suitcase: $P_{\text{contr}} = 1$; two suitcases: $P_{\text{contr}} = 0.5$, 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.

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 Héту, 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).

- Shen FX, Hoffman MB, Jones OD, Greene JD, Marois R (2011) Sorting guilty minds. *NYU Law Rev* 86(5):1306–1360.
- 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.
- Haynes JD, Rees G (2006) Decoding mental states from brain activity in humans. *Nat Rev Neurosci* 7(7):523–534.
- Hughes V (2010) Science in court: Head case. *Nature* 464(7287):340–342.
- Treadway MT, et al. (2014) Corticolimbic gating of emotion-driven punishment. *Nat Neurosci* 17(9):1270–1275.
- 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.
- 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.
- Preuschoff K, Bossaerts P, Quartz SR (2006) Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 51(3):381–390.
- Preuschoff K, Quartz SR, Bossaerts P (2008) Human insula activation reflects risk prediction errors as well as risk. *J Neurosci* 28(11):2745–2752.
- 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.
- 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.
- d'Acremont M, Schultz W, Bossaerts P (2013) The human brain encodes event frequencies while forming subjective beliefs. *J Neurosci* 33(26):10887–10897.
- McCoy AN, Platt ML (2005) Risk-sensitive neurons in macaque posterior cingulate cortex. *Nat Neurosci* 8(9):1220–1227.
- Ahn WY, et al. (2014) Nonpolitical images evoke neural predictors of political ideology. *Curr Biol* 24(22):2693–2699.
- Ryali S, Supekar K, Abrams DA, Menon V (2010) Sparse logistic regression for whole-brain classification of fMRI data. *Neuroimage* 51(2):752–764.
- Haxby JV, Connolly AC, Guntupalli JS (2014) Decoding neural representational spaces using multivariate pattern analysis. *Annu Rev Neurosci* 37:435–456.
- Gabrieli JD, Ghosh SS, Whitfield-Gabrieli S (2015) Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* 85(11):11–26.
- Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. *Nat Neurosci* 8(5):679–685.
- 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.
- Tong F, Pratte MS (2012) Decoding patterns of human brain activity. *Annu Rev Psychol* 63:483–509.
- Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc B* 67:301–320.
- Whelan R, Garavan H (2014) When optimism hurts: Inflated predictions in psychiatric neuroimaging. *Biol Psychiatry* 75(9):746–748.
- 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.
- Singer T, Critchley HD, Preuschoff K (2009) A common role of insula in feelings, empathy and uncertainty. *Trends Cogn Sci* 13(8):334–340.
- Miller EK (2000) The prefrontal cortex and cognitive control. *Nat Rev Neurosci* 1(1):59–65.
- 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.
- Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481):453–458.
- 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.
- 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.
- Henrich J, Heine SJ, Norenzayan A (2010) The weirdest people in the world? *Behav Brain Sci* 33(2-3):61–83, discussion 83–135.
- 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.