Detecting *Mens Rea* in the Brain

169 U. PENNSYLVANIA LAW REVIEW (forthcoming 2019)

[Version of April 29, 2020]
[In Progress; Do Not Cite Without Permission]

by

Owen D. Jones
Glenn M. Weaver, M.D. and Mary Ellen Weaver Chair in Law, Brain, and Behavior
Professor of Law & Professor of Biological Sciences
Director, MacArthur Foundation Research Network on Law and Neuroscience
Vanderbilt University

Read Montague
Vernon Mountcastle Research Professor
Professor of Physics
Director, Human Neuroimaging Lab and The Computational Psychiatry Unit
Virginia Tech
Member: *MacArthur Foundation Research Network on Law and Neuroscience*

Gideon Yaffe
Wesley Newcomb Hohfeld Chair of Jurisprudence
Professor of Philosophy & Professor of Psychology
Yale University
Member: *MacArthur Foundation Research Network on Law and Neuroscience*

Abstract

What if the widely used Model Penal Code (MPC) assumes a distinction between mental states that doesn't actually exist? The MPC assumes, for instance, that there is a real distinction in real people between the mental states it defines as "knowing" and "reckless." But is there?

If there are such psychological differences, there must also be brain differences. Consequently, the moral legitimacy of the Model Penal Code's taxonomy of culpable mental states – which punishes those in defined mental states differently – depends on whether those mental states actually correspond to different brain states in the way the MPC categorization assumes.

We combined advanced functional brain-imaging technology with new artificial intelligence tools to see if the brain activities during *knowing* and *reckless* states of mind can ever be reliably distinguished.

As our experiment indicates, the answer is Yes. So here we provide an overview of our brain-scanning experiment, discuss important implications, and detail several necessary precautions, so our results won't be over- or mis-interpreted.

Table of Contents

Detecting *Mens Rea* in the Brain

Owen D. Jones, Read Montague, and Gideon Yaffe[1]

Introduction

Mental states matter. Consequently, we and colleagues designed and executed a brain-imaging experiment attempting to detect – for the first time – differences between mental states relevant to criminal law.

To set this up, let's suppose that you've just killed someone in Colorado. It was not your purpose or desire to kill him. Nevertheless, another human being is dead. Arrested and on trial, you do not dispute that your action unjustifiably caused his death. But whereas the prosecutor argues that you *knew* someone would die as an inevitable by-product of your actions, you assert in your defense that you knew no such thing. Instead (you claim) you were merely *reckless*. That is, you acted as you did with awareness of a substantial risk that someone would be fatally injured, but without any express purpose to kill anyone.

Now it turns out (and this part isn't hypothetical) that in Colorado, as in many states, there is a huge difference in the ranges of sentencing outcomes between being convicted of a *knowing* homicide and a *reckless* one. In Colorado it

---

[1] Jones holds the Glenn M. Weaver, M.D. and Mary Ellen Weaver Chair in Law, Brain, and Behavior at Vanderbilt University, where he is Professor of Law, Professor of Biological Sciences, and Director of the *MacArthur Foundation Research Network on Law and Neuroscience*.

Montague holds the Vernon Mountcastle Research Professorship at the Fralin Biomedical Research Institute at Virginia Tech, where he is also a professor in the department of Physics, director of the Human Neuroimaging Lab and The Computational Psychiatry Unit, and is a member of the *MacArthur Foundation Research Network on Law and Neuroscience*.

Yaffe holds the Wesley Newcomb Hohfeld Chair of Jurisprudence at Yale Law School. He is also Professor of Philosophy and Psychology at Yale, in addition to being a Member of the *MacArthur Foundation Research Network on Law and Neuroscience*.

We wish to acknowledge our terrific colleagues on the brain-scanning experiment described here. The data collection and the analysis of the data were done in P. Read Montague's lab at Virginia Tech by Iris Vilares, Michael J. Wesley, Woo-Young Ahn, Terry Lohrenz, and Montague. Richard J. Bonnie, Morris Hoffman, and Stephen J. Morse played an important role in the design of the experiment and advised the project throughout. The study was supported by a grant from the John D. and Catherine T. MacArthur Foundation to Vanderbilt University, with a subcontract to Virginia Tech. Support for the present article was provided, in part, by the MacArthur Foundation and the Glenn M. Weaver Foundation. This article does not necessarily represent official views of either the MacArthur Foundation, the MacArthur Foundation Research Network on Law and Neuroscience, or the Weaver Foundation. Michael Dunbar provided helpful research assistance.

means the difference between being sentenced to 16-to-48 years in prison *and none*.[2] So your fate rests in the hands of lay jurors who will decide what your mental state was at the time of the fatal act. Specifically: Did you *know* you would kill someone, or were you merely aware of a *risk* that you would? And here you face two very large problems, both of which our current criminal justice system ignores.

First, the legal system assumes that jurors can reliably distinguish, in the way the legal system contemplates, between the two mental states at issue. It is well known to all first-year law students that the supermajority of states follow the Model Penal Code's (MPC) long-standing approach to categorizing culpable mental states into four types: purposeful, knowing, reckless, and negligent.[3] Large numbers of offenses, including homicides, are then subdivided into corresponding categories. And punishment severity follows accordingly. But less well known is that a body of experimental evidence suggests that jurors are not really all that good at understanding which category is which.[4] Subjects frequently get it wrong, even when directly instructed on the relevant legal definitions and standards. And they find it particularly difficult to sort defendants between the MPC's categories of "knowing" and "reckless." They confuse the two about 50% of the time, under

---

[2] Second degree murder, without any heat of passion mitigator, is defined and classified as a Class 2 felony at COLO. REV. STAT. §§ 18-3-103(1), -103(3)(a) (2010). Class 2 felonies ordinarily carry a non-mandatory presumptive sentence of eight to twenty-four years. *Id.* § 18-1.3-401(1)(a)(V)(A). However, murder is often considered to be a crime of violence, a determination that has the effects of (1) increasing the range to sixteen to forty-eight years; and (2) making a prison sentence mandatory. *Id.* § 18-1.3-406 (pertaining to murders involving deadly weapons or to crimes causing serious bodily harm or death). By contrast, a reckless murder is classified as manslaughter, and caries a non-mandatory sentence of two to six years. Manslaughter is defined and classified as a Class 4 felony in Colorado. Id. § 18-3-104. Class 4 felonies carry a *non-mandatory* presumptive sentence of between two and six years. Id. § 18-1.3-401(1)(a)(V)(A.1). Manslaughter is not defined as a crime of violence under § 18-1.3-406.

[3] In addition, even judges in jurisdictions that have not adopted Model Penal Code language in defining mens rea categories are deeply influenced by it in their interpretations of many statutes. The reach of the MPC's mens rea regime, that is, extends beyond those states that have explicitly adopted it.

[4] *See* Francis X. Shen, Morris B. Hoffman, Owen D. Jones, Joshua D. Greene, and Rene Marois, *Sorting Guilty Minds*, 86 N.Y.U. L. REV. 1306 (2011) [hereafter *Sorting Guilty Minds*]; Matthew R. Ginther, Francis X. Shen, Richard J. Bonnie, Morris B. Hoffman, Owen D. Jones, Rene Marois, and Kenneth W. Simons, *The Language of Mens Rea*, 67 VANDERBILT L. REV. 1327 (2014) [hereafter *Language of Mens Rea*]; Matthew R. Ginther, Francis X. Shen, Richard J. Bonnie, Morris B. Hoffman, Owen D. Jones, Kenneth W. Simons, *Decoding Guilty Minds*, 71 VANDERBILT L. REV. 241 (2018) [hereafter *Decoding Guilty Minds*]. *See also*, Kevin John Heller, *The Cognitive Psychology of Mens Rea*, 99 J. CRIM. L. & CRIMINOLOGY 317 (2009); James McLeod, *Belief States in Criminal Law*, 68 OKLAHOMA L. REV. 497 (2016); Justin Levinson, *Mentally Misguided: How State of Mind Inquiries Ignore Psychological Reality and Overlook Cultural Differences*, 49 HOWARD L. J. 1 (2005).

some conditions, and do little better under others.[5] In such a case, that's nearly coin-flipping odds of an incorrect conviction. So that's one very big problem with the current system.

The other big problem – the elephant in the room, in fact – is that despite the legal system's bald assumption that these two mental states are different, no one even knows if the legally assumed and statutorily instantiated distinction between knowing and reckless mental states reflects an actual and inherent psychological difference. The supposed distinction between them could be nothing more than a convenient fiction, upon which countless trials – and far more plea bargains – have been built.

To see the even deeper bite of this second problem, first consider the way in which Legal Realists have approached the mens rea categories. The Legal Realist tradition is well-known for the claim that many legal terms and concepts falsely purport to classify defendants and their circumstances on the basis of their intrinsic features.[6] Instead, assert the Realists, defendants are classified on the basis only of the judge's desire to hold some liable and to decline to hold others liable,

---

[5] *Sorting Guilty Minds*, *supra* note xx, at 1351. This particularly troubling finding helped prompt a number of additional experiments, published in *The Language of Mens Rea*, *supra* note xx, and *Decoding Guilty Minds*, *supra* note x. Note that subjects are typically less able to correctly identify knowing and reckless scenarios than to correctly identify the other mental states, even when given definitions of the mental states. *Sorting Guilty Minds*, supra note xx, at 1348. Even when the ability to classify correctly is improved under experimental conditions, using variations on definition language, knowing and reckless mental states remain by far the hardest to classify. *Language of Mens Rea*, *supra* note xx, at xx. Indeed, under such circumstances, even in the best case only 59% of subjects accurately identify reckless scenarios as reckless. And 70% of those misidentifications confuse a reckless scenario for a knowing one. *Id.* at 1356. Together, these error rates mean that subjects presented with a reckless scenario categorized it as a knowing scenario a whopping 41% of the time.

A further difficulty is that even those subjects who correctly classify knowing and reckless scenarios 75% of the time quite often do not rank the knowing ones as more punishment-worthy than the reckless ones. *Sorting Guilty Minds*, *supra* note xx, at 1344. And the failure consistently to draw rank-ordered distinctions between the two culpable mental states in the way the MPC does appears to hold even when subjects are instructed on how the two mental states are defined (id., at 1339-41), and even when those instructions are provided with clearer definitions, and related variants. *Language of Mens Rea*, *supra* note xx, at 1351-53. This raises important questions about the normative basis for the knowing/reckless distinction that are, though crucial, beyond the scope of this article.

[6] The Realists were especially known for offering this critique of legal concepts like "causation" and "corporation." Whether the defendant "proximately caused" the plaintiff's harm, Legal Realists argue, turns not on the presence or absence of any liability-independent features of the case— such as the "reasonable foreseeability" of the harm, or the absence of "voluntary intervention." Rather, judges *claim* to be deciding cases on the basis of such features when what really decides the question is something else, something about the judge or his or her views about what makes for sound policy. See, for instance, Felix Cohen, *Transcendental Nonsense and the Functional Approach*, 35 COLUMBIA. L. REV. 809 (1935)*;* Karl Llewellyn, THE BRAMBLE BUSH (1930).

even when those two groups of people do not differ in any way other than in the eye of the judge.[7] But then note that this critique has extended to juries as well. That is, the Legal Realists claim that the question of whether you were knowing or reckless when you performed the act that killed someone only *purports* to be a question about your psychology.

The important, long-lingering question is therefore: Does the distinction between MPC mens rea categories, such as knowing and reckless, reflect an intrinsic psychological difference, actually found in human beings? If so, we believe that one should expect in principle that there would also be a difference between the brains of reckless and knowing individuals, at the times of their actions. Because, after all (and setting aside some philosophical subtleties[8]) anytime there is a psychological difference there must also be a brain difference.

So is there a neural difference or not? Lives and liberties ride on the answer to that simple, straightforward, and pointed question for thousands each year who stand accused before the criminal justice system. For whenever we have used the supposed distinction between knowing and reckless (and other MPC categories) to justify a different punishment under the law, when there is in fact no detectable or meaningful psychological distinction, then widespread injustice will have followed in the wake of the MPC, and will continue indefinitely, if unchecked.

As it turns out, there's a good reason why the supposed distinction in the brain between those who are knowing and those who are reckless has never been tested empirically. There's simply never been a good way to investigate whether there are or are not any discernible differences in the brain activity of people in

[7] Dan Kahan's two notable papers on mistakes in criminal law are naturally construed as offering just this kind of critique of *mens rea* concepts. Dan M. Kahan, *Ignorance of the Law is an Excuse – But Only for the Virtuous*, 96 MICH. L. REV. 127 (1997); Dan M. Kahan, *Is Ignorance of Fact an Excuse Only for the Virtuous?*, 96 MICH. L. REV. 2123 (1998). According to Kahan, the law allows people with no relevant intrinsic psychological differences to be distinctly classified as having made, or having failed to make, an exculpatory mistake. *See also* Thurman W. Arnold, *Criminal Attempts – The Rise and Fall of an Abstraction*, 40 YALE L.J. 53, 68-9 (1930) (arguing the different concepts of "intent" are *post fact* ways of rationalizing verdicts reached for independent reasons); Janice Nadler, *Blaming As A Social Process: The Influence of Character and Moral Emotion on Blame*, 75 LAW & CONTEMP. PROBS. 1 (2012) at 4 ("[M]oral character might serve as a kind of proxy for mental state, so that a person with a bad character is blamed as if he were reckless, whereas a person with a good character is blamed as if he were not reckless.").

[8] The philosophical literature concerned with the view labeled "externalism about mental content" concerns the possibility that mental states could vary even without variation in brain activity, and without postulating the existence of some non-material aspect to mind. The classic statement of the view is found in Hilary Putnam, *The Meaning of Meaning in* PHILOSOPHICAL PAPERS, VOL. II: MIND, LANGUAGE, AND REALITY (1975). A useful overview of the current state of the literature is Joe Lau and Max Deutsch *Externalism About Mental Content*, THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY, https://plato.stanford.edu/entries/content-externalism/.

these allegedly different mental states. The putative psychological differences were too subtle to be investigated using the technology we had.

Until now.

With a grant of nearly $600,000 from the *MacArthur Foundation Research Network on Law and Neuroscience*,[9] we – as part of a larger interdisciplinary team[10] – set out to investigate the knowing-reckless distinction in the brain, and the boundary that may separate them. Specifically, we set out to see if we could use brain activity alone to detect the difference between those who the law would classify as "knowing" and as "reckless." By combining the relatively new technical achievements of functional magnetic resonance imaging (fMRI) with new advances in the analytic abilities of machine-learning algorithms (a form of artificial intelligence) our team conducted the first ever assault on this thorny legal problem.

This article for the first time reports and describes, for a legal audience, the results and implications of our experiment. Cutting to the chase: we found evidence strongly supporting the existence of a brain-based distinction between knowing and reckless mental states. Our detailed neuroscience paper was first published (as rules of scientific publications demand) in a dedicated peer-reviewed science journal, the *Proceedings of the National Academy of Sciences*.[11] It received some sensationalist press coverage, including headlines such as this one from the British *Daily Mail:* "Something On Your Mind? AI Can Read Your Thoughts and Tell Whether You are Committing a Crime."[12] Half truths like these are dangerous. So our goal here is to explain for a legal, non-scientific audience what we did and – more importantly – how it does and does not matter for the law.

Our team's discovery is relevant to law in two ways.[13] First, it provides new information, of an entirely novel and cutting-edge kind, relevant to the

---

[9] One of us (Jones) designed and directs the MacArthur Foundation *Research Network on Law and Neuroscience*, which is headquartered at Vanderbilt University and funded by over $7,500,000 in grants from the John D. and Catherine T. MacArthur Foundation. The *Research Network* partners selected leading legal scholars, neuroscientists, and judges from around the country for intensive collaborative work on law-relevant neuroscience experiments. For further information on the *Research Network*, its activities, and its more than 85 publications, see its website at www.lawneuro.org.

[10] One of us (Yaffe) led the *Working Group on Detection and Classification* in collaboration with neuroscientist Read Montague. The full interdisciplinary team, in alphabetical order, consisted of: Woo-Young Ahn, Richard J. Bonnie, Morris B. Hoffman, Owen Jones, Terry Lohrenz, Read Montague, Stephen Morse, Iris Vilares, Michael Wesley, Gideon Yaffe.

[11] Iris Vilares, Michael Wesley, Woo-Young Ahn, Richard J. Bonnie, Morris B. Hoffman, Owen D. Jones, Stephen J. Morse, Gideon Yaffe, Terry Lohrenz, & Read Montague, *Predicting the Knowledge-Recklessness Distinction in the Human Brain*, 14 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 3222 (2017).

[12] March 13, 2017.

[13] For an overview of ways neuroscience can be relevant to law, see Owen D. Jones, *Seven Ways Neuroscience Aids Law*, in NEUROSCIENCES AND THE HUMAN PERSON: NEW PERSPECTIVES ON

substantive debates over the accuracy and legitimacy of the distinction the Model Penal Code draws between knowing and reckless mental states. In this one important domain, that is, there is reason to think that what the law purports to do—draw a distinction in liability on the basis of a distinction in psychological state—is what it actually does. Second and collaterally, our results serve as a clear, concrete, and salient example of how new neuroscientific techniques, sometimes partnered (as here) with artificial intelligence tools, can be used to probe matters of legal relevance.

We proceed in three primary parts. Part I provides an overview of the experiment and the results. Along the way, it offers a necessary but brief and accessible introduction to how fMRI brain imaging works. Part II discusses the important implications and promise of this new finding – as well as, correlatively, the non-implications. Part III provides necessary caveats and cautions, to ensure that our results won't be over- or mis-interpreted.

Part I: Detecting *Mens Rea* in the Brain

A. Background: Initial Obstacles

The biggest challenge for our team was to develop an experimental paradigm that could elicit states of mind that legal scholars and the legal system in general would consistently classify as knowing and reckless. It was obvious, at the outset, that we couldn't actually scan the brains of criminals in the midst of criminal behavior. Almost all brain scanning techniques require a willing participant, after all. And most require that the participant stay motionless for an extended period (often as much as 50 minutes).

It was also obvious that – even with willing and stationary lab subjects – we couldn't have them actually commit any of the few criminal acts one could theoretically commit while being scanned, such as computer hacking to steal data or funds. For one thing, no institutional review board (which signs off on the ethics of human experiments) would allow us to direct or encourage subjects to engage in criminal behavior. For another, the process of hacking takes too long, and would have too many uncontrolled variables between subjects, to enable useful scanning. And that's even assuming we could find a set of subjects with the needed skills.

So after a lengthy process of brainstorming, false starts, and investigative work over a roughly two-year period, amongst ourselves and the broader group of co-Members of the *MacArthur Foundation Research Network on Law and*

HUMAN ACTIVITIES (A. Battro, S. Dehaene & W. Singer, eds., Scripta Varia: Pontifical Academy of Sciences, 2013). *See also* Owen D. Jones & Anthony D. Wagner, *Law and Neuroscience: Progress, Promise, and Pitfalls*, in THE COGNITIVE NEUROSCIENCES (Gazzaniga, Mangun, and Poeppel, eds., 6th edition, forthcoming 2019) (connecting those ways to experimental results).

Neuroscience, we ultimately settled on a promising paradigm. Initially, our goal had been to see whether we could determine, based solely on information about a person's brain activity, whether her psychological state was properly classified in one rather than another of the Model Penal Code's four *mens rea* categories. That is, we were not focused, at first, on the knowing-reckless distinction in particular; we also developed designs for experiments aimed at drawing the purposeful-knowing distinction, and the reckless-negligent distinction.

Knowing that we could not pursue all avenues at once, we bet on an experiment that we believed could get at the root of the knowing-reckless distinction, given how the MPC defines these supposedly different mental states, thereby providing a proof of concept for brain imaging mens rea. At its core, the design has subjects make repeated decisions about whether to take risks, when the risks vary from one scenario (trial) to the next, and with differing gains and losses accompanying the different risks.

We already provided, in the Introduction, a reminder of how the knowledge-recklessness distinction plays out in the context of homicide. To transition from that context to our experimental paradigm, let us now provide another example, in the context of theft.

Suppose two different scenarios in which a person takes something that does not belong to him without permission. In both scenarios, our defendant emails his neighbor to ask permission to borrow that neighbor's car. Assume further that, in the past, the neighbor has said yes to this request about half the time, and about half the time has said no.

In the first scenario, our defendant checks his email and sees that the neighbor has said no. Then the defendant borrows the neighbor's car anyway, knowing full well that he does not have permission to do so. He figures (incorrectly as it turns out) that the neighbor won't even notice.

In the second scenario, our defendant never checks his email for a reply, and therefore never sees that the neighbor has said no. Then the defendant goes ahead and borrows the car anyway. He figures there's about a fifty percent chance he has permission, and he also figures (incorrectly) that either way the neighbor will probably never notice.

The MPC would classify the first defendant as liable for a knowing theft and the second as liable for a reckless theft. So, a central component of the distinction between knowing and reckless mental states, under the MPC regime, is that a person in a knowing state of mind is essentially 100% certain about the presence of an element of a crime. In contrast, a person in a reckless state of mind can have a belief about the probability located within a range—not so low as to promise a *de minimis* expected harm, but not so high as to be functionally equivalent to certainty. That range –defined imprecisely under the Model Penal Code as encompassing a "substantial and unjustifiable risk"—is to some degree

context sensitive: a 10% chance of killing someone if you act will certainly qualify as a "substantial and unjustifiable risk" to impose, while a 10% chance that you might trample his tomato plants might not. So what we needed was an experiment in which subjects would sometimes choose to perform an act while knowing a certain condition was in place, and sometimes choose to perform the same act while aware, instead, of a risk high enough to potentially qualify as "substantial and unjustifiable" while still far enough below 100% as to fall short of knowledge.

B. The Paradigm: Eliciting Knowing and Reckless Mental States

We asked our subjects to imagine that they were given an opportunity to carry a briefcase across the border.[14] The briefcase might or might not contain "valuable content" (such as documents or microchip processors).[15] And a carried briefcase might or might not be searched at the border.

There was a significant financial incentive to choose to carry the briefcase. Specifically, subjects who could get a briefcase across the border, without being searched, could leave the lab with quite a lot of money. But getting caught at the border carrying the contraband resulted in a financial penalty. The other two options—getting "caught" with an empty briefcase, or crossing successfully with an empty briefcase—had payoffs in between.

What we were primarily interested in was the differences between our subjects' brain activity when they were certain that the briefcase contained contraband and their brain activity when there was some "substantial and unjustifiable" risk that it did. While subjects made decisions about whether to carry briefcases across the border, we scanned their brains and collected data.

We instructed subjects before the game began on the details of the payoff structure. (See Figure 1.) Subjects began each of the 125 trials they completed with a hypothetical $6,000 in the bank. The payoff structure then governed how much a subject could earn or lose from the intersection of her choice (carry or don't carry) with two variables: 1) the probability that a carried case contained contraband (the "Contraband Risk"); and 2) the probability that a carried case would be searched (the "Search Risk").

[14] The subjects for this experiment were 40 in number, half of them female, half of them male. Their average age was about 29.

[15] We used the phrase "valuable content" with subjects instead of, say, "illegal drugs" to reduce the possibility that some subjects would always refuse the option to carry a case, on moral or legal grounds. (None did.) For brevity, however, we will hereafter refer to the valuable content as "contraband."

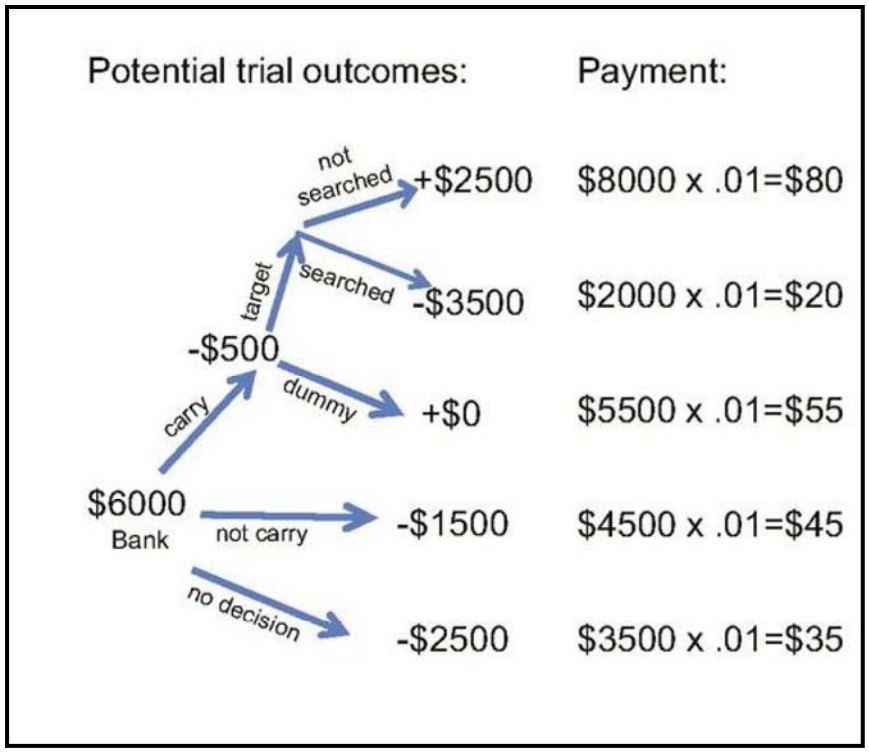

Figure 1. Payoff Structure

Specifically, a subject gained $2,000 (in her virtual bank account) each time she carried a case containing contraband through a checkpoint unapprehended. But she lost $4000 from that account if she carried a case with contraband and got caught. If she carried a case that turned out not to contain contraband at all, regardless of whether or not her case was searched, she lost $500. To give our subjects incentive to make a choice, and an incentive to choose in favor of carrying, we added two more fees: choosing not to carry cost $1,500, and refusing to choose whether or not to carry cost $2,500. We also informed subjects that if they completed the experiment they would receive, in cash to take home, one percent of the payout from one trial, chosen at random. Given the details of the payoff structure, each subject therefore received between $20.00 and $80.00 for participating. These payoff details are illustrated in Figure 1.

Now to the in-trial details. At the beginning of a trial, we provided each subject with two pieces of information. We informed them of the Contraband Risk and the Search Risk. To communicate Contraband Risk, we showed the subject a picture of one, two, three, four, or five cases – only one of which ever contained contraband. We informed subjects that: a) this was the pool of cases from which

the subject's own case would be chosen; and b) one and only one briefcase in that pool contained contraband. See Figure 2.

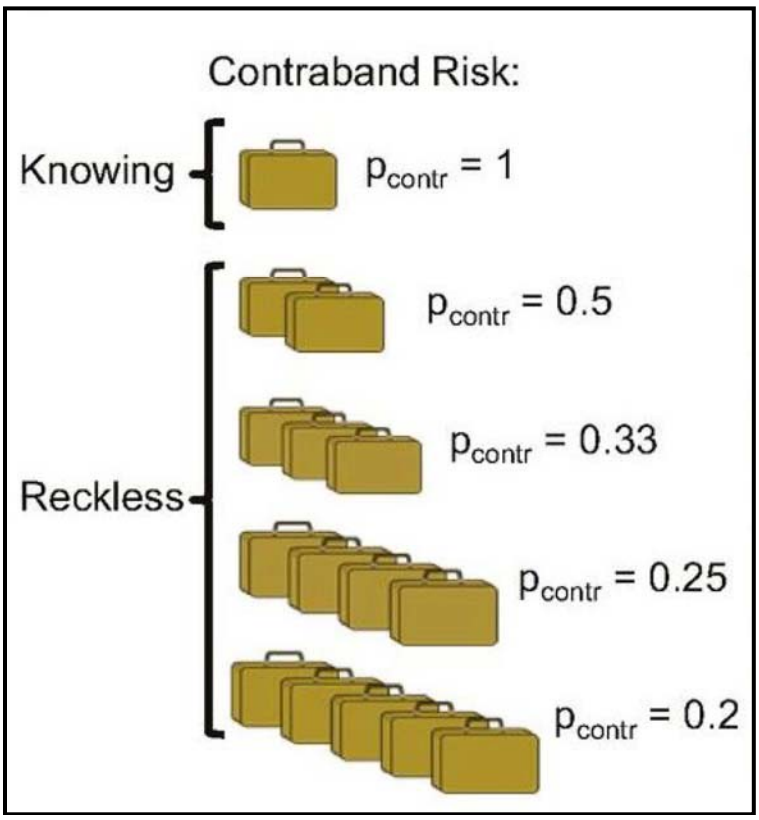

Figure 2. Contraband Risk
"Pcontra" = Probability of Carrying Contraband (varying from 20% to 100%)

Thus, if the subject saw a picture of five cases that subject could deduce that choosing to carry a randomly assigned case in that trial yields one-chance-in-five (or a 20% likelihood) of getting a case with contraband. Conversely, if the subject saw only one case, and understood that only one case would contain contraband, that subject could deduce that choosing to carry that case yields a one-chance-in-one (or 100% likelihood) of getting a case with contraband. And so on for two, three, and four cases.

To signal the probability that a carried case would be searched, we showed each subject, on each trial, a picture of ten tunnel exits, some number of which – either two, four, six, or eight -- showed a guard standing prominently in the exit. See Figure 3. As with cases, subjects could readily calculate that the probability of

being searched if there were eight guards was much higher (eight-chances-in-ten, or 80%) than if there were only two guards (two-chances-in-ten, or 20%).

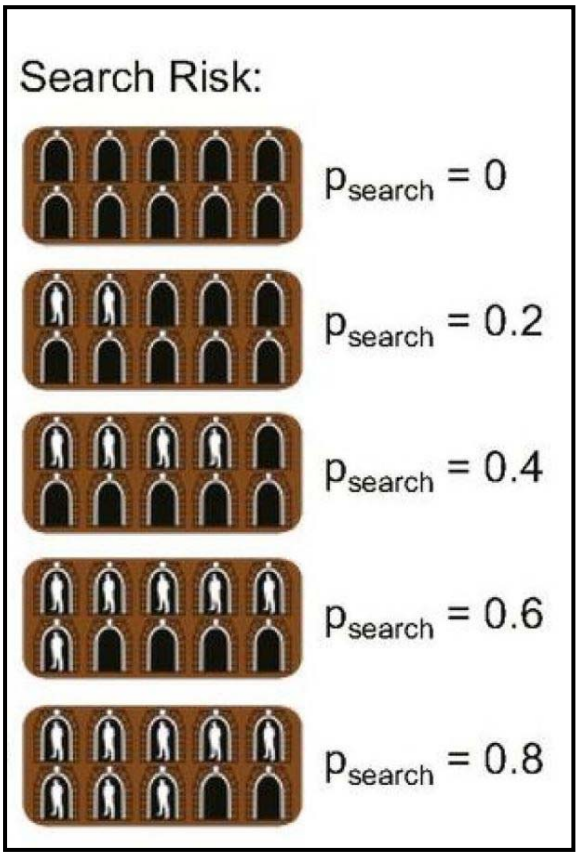

Figure 3. Search Risk
"Psearch" = Probability of Search (varying from 0% to 80%)

As a standard precaution, and to counterbalance any effect on our subjects of the order in which we showed the two images that signal these two pieces of information – Contraband Risk and Search Risk, respectively -- half of our subjects always learned of the Contraband Risk before they learned the Search Risk. And the other half of the subjects always learned the Search Risk before they learned the Contraband Risk. As we'll see below, this turned out to be important.

After learning these two pieces of information, and mindful of the payoff structure, a subject was tasked to indicate whether she would be willing to carry a case (selected at random from the cases presented) through a tunnel selected at random from among those 10 tunnels presented. After subjects registered their

choices to carry or not carry, there was a brief pause before the subjects would start over again in a new trial, with a new allotment of hypothetical money, new information, and a new chance to carry or not carry a case.

Importantly, subjects never learned whether, for a given trial, a case they carried was searched. Similarly, they never learned whether the case they carried did or did not contain contraband. This was important because we did not want our subjects' decisions in any given trial to be influenced by the results of the previous trial. That is, we did not want our subjects to make inferences of the form, "On the last trial, I was carrying an empty case, so I bet I get one with contraband this time." We wanted each trial to be as close to a one-shot decision in the face of risk as we could engineer.

C. Virtues of the Paradigm

There are several virtues to the design. First, the paradigm can draw a clear line between subjects who are and aren't in a "knowing," as distinct from "reckless," mental state. Starting with the "knowing" condition, recall that we informed subjects that one and only one case will ever contain contraband. For this reason, we can reasonably believe (absent inconsistent behavior to the contrary) that when our subject chooses to carry the single case offered (i.e., a case that *must* contain contraband) she is in a "knowing" mental state with respect to the case containing contraband. In contrast, any time she encounters the choice to carry a random case selected from either two, three, four, or five presented cases, and also chooses to carry, we can reasonably believe that she is aware of the respectively varying degrees of probability that she is carrying the contraband, and is therefore in a "reckless" state of mind.

Second, because we did not inform our subjects directly of the risks—e.g. by describing those probabilities as "50%" or "1-in-2"—we mimicked an important feature of many real cases. Specifically, people in real world situations ordinarily infer probabilities from evidence, rather than being presented with numeric summary information about probabilities. For instance, when someone is deciding whether to run the red light, there is no sign hanging in the air that says, "The probability of killing someone by doing that is 19%." Rather, one reaches a judgment about the probability by looking at the number of oncoming cars, their speed, etc. In the experiment, our subjects had to infer the two relevant probabilities from a picture of that round's pool of cases, and a picture of tunnels (some fraction of which had guards in them, that round). This is not, of course, the form that most evidence of probability takes in real life. But it is far closer than would be directly presented numerical information.

Third, varying the chance of being caught from 20% to 80% further allowed us to mimic another feature of real cases: people who commit crimes often decide

to do so in part by calculating the risks of being apprehended. Juries assessing mens rea, however, are never asked to determine what probability the defendant assigned to his being caught; that's not relevant to the mens rea inquiry. What matters is what probability (in lay, rather than statistical terms) the defendant assigned to legally relevant elements of the crime, such as killing someone else, or not having permission to borrow, or there being drugs in the briefcase he was carrying. Controlling the information about the chance of detection, boosted our ability to do what fact-finders are asked to do: meaningfully distinguish between the awareness of the risk that is relevant to a recklessness assessment (namely the awareness of the risk that the case contained contraband) from the awareness of the risk of apprehension, which is not relevant to the question of recklessness.

Fourth, varying the pool of cases from one to five created the possibility that we might learn something about how brain states vary within the reckless mental state itself. That is, they might vary as a function of the changing probabilities that one would be carrying contraband – from 20% (if five cases were presented) to 50% (if only two cases were presented).

There are some important limitations to this laboratory experiment, as there are with any experiment. We will be explaining those in Part III, below. But at this point the key thing to see, about the experimental paradigm, is that it provides us with a range of decisional processes and behavioral outputs that vary in two ways. First, between knowing and not knowing that one will be carrying contraband (if one chooses to carry at all). Second, between four degrees of not knowing, all of which can be characterized as reflecting varying quanta of recklessness.

The core idea is that collecting data on brain activity during each trial, and analyzing that data in conjunction with the varying behavioral outputs (i.e., choosing, each trial, whether or not to carry), should afford us some window on whether, and if so how, neural activity varies between knowing and reckless conditions.

D. Tools for Detecting *Mens Rea*

But how, exactly, did we collect and analyze the data? In this subsection we provide a brief overview of how fMRI brain-imaging works, and how machine-learning algorithms assist in finding useful and predictive patterns in the data. Readers familiar with these can, of course, skip to the next section.

1) *fMRI Brain Imaging, Generally*

Prior to the invention of functional magnetic resonance imaging (fMRI) in the early 1990s, researchers had sophisticated tools for measuring the physical

structure of the brain, but somewhat limited tools for measuring brain *function*. Techniques for studying brain structure then, as now, include not only direct dissection of the brain after death but also various brain-imaging tools such as x-rays, computed tomography scans (CT), and magnetic resonance imaging scans (MRI). But until the early 1990s our ability to measure how neurons (the cells that make up the brain) were behaving while people were performing tasks of various sorts was limited to electroencephalography (EEG) (which measures electrical activity at the scalp caused by neural activity in the brain) and positron emission tomography (PET) (which measures the behavior in the brain of radioactive molecules injected for the purpose). Although both are powerful tools, useful for many things, they are also quite limited. EEG gives us imperfect information about what's happening in the interior structures of the brain, far from the scalp. And PET, since it involves expensive and invasive injections, cannot be used to study the brains of large numbers of healthy people. Although it is also possible to directly record electrical activity in the brain from electrodes implanted into brain tissue, that is only possible in humans who are already undergoing brain surgery, or in animals, whose brains may or may not function much like those of people. Put another way, we could at the time get very clear measures of your heart as it pumps blood, and of your muscles as they contract, but we had limited ways of measuring your brain while you were using it.

fMRI significantly changed all that. And that's because: a) it enables strong inferences about neural activity within and across the entire brain; and b) it is sufficiently noninvasive that it can be used on healthy people without surgery or injection of any sort.

Over the last 25 years, fMRI has become one of the world's most dominant research tools for learning about brain function. And although its details are both technical and elegant (and there are many technical descriptions available for the most motivated reader[16]) we provide here a brief and accessible overview, so that readers may understand the nature of the brain-imaging experiment that we and our colleagues conducted. We will start with the big picture, and drill down just far enough – with a minimum of technical jargon – to provide a basic appreciation of how this works.

At the big picture level, you can analogize the fMRI process to a bat's echolocation. In the same way that a bat sends a wide high-frequency sound at small potential targets, and then makes strong inferences about their locations from the directions of sound reflected back, fMRI beams radiowaves to the brain, and enables inferences from the differential patterns in energy that returns from within

[16] *See, e.g.,* Scott A. Huettel, Allen W. Song & Gregory McCarthy, FUNCTIONAL MAGNETIC RESONANCE IMAGING (3rd ed. 2014); Robert W. Brown, Y.-C. Norman Cheng, E. Mark Haacke, Michael R. Thompson, & Ramesh Venkatesan, MAGNETIC RESONANCE IMAGING: PHYSICAL PRINCIPLES AND SEQUENCE DESIGN (2nd ed. 2014).

brain tissues. More specifically, fMRI enables researchers to discover and monitor both the locations of changes in blood flow, and the amounts of those changes, correlated with the different moments in each subject's information-gathering, information-processing, and decision-making tasks, as well as with the decision itself.

Here's a little more context. Researchers place a subject on her back within a large tube that is surrounded by massive, super-cooled, super-conducting wire coils arranged to move electrical energy in a particular pattern. Specifically, the coils are arranged to create a very strong magnetic field, within the scanner, that can be exquisitely manipulated (and even graduated in strength, along an axis) in each of the three dimensions – length, width, and height.

There are then several additional things to know, before turning to explore how this works. First, all atoms (including those in the body) contain some spinning particles, each bearing an electrical charge. Second, spinning objects with an electrical charge are, in themselves, tiny magnets. Third, placing a person within a strong magnetic field of an MRI tends to align the axes of spin of their subatomic particles, just as metal filings on paper will align with a field of a magnet held underneath.

Let's step back and start connecting this to neurons in the brain. Neurons are the brain cells – part of the nervous system – that carry electrical impulses from one end to another and that, by virtue of their interactions with other brain cells, enable everything from perception to decision to action. Like all cells, they need nutrients supplied by the blood – such as oxygen – in order to live and function. The more active neurons are, the more oxygenated blood they need.

Which brings us to the happy fortuity that enables fMRI to discover things about brain function: Oxygenated blood cells (bringing oxygen to the neurons) and de-oxygenated blood cells (which have already off-loaded their oxygen to neurons) have different magnetic properties. The significance is this. When an MRI "pings" (so to speak) a brain in the scanner, certain subatomic particles that are all spinning in the same axis are temporarily bumped out of alignment. And when the signal stops, and those subatomic particles snap back into alignment with the magnetic field, they release a certain amount of energy, which can be spatially located, in the brain, by an array of receivers in the MRI machine.

Because fMRI technology can detect changing ratios in oxygenated and deoxygenated blood, over both time and space, researchers can make inferences about where different brain regions are most and least active, during each of the trials a subject undertakes. Researchers then compare that information either to a baseline of brain activity (the so-called "resting state", when an awake brain is simply talking quietly to itself, without any specific task to perform, other than normal bodily functions) or to a contrasting set of decisions that task the brain in a

different way. And this enables them to learn about how the brain operated during the particular decisions they are studying.

Put another way, just as a bat can place a mosquito in airspace, on the basis of reflected sound waves, fMRI can detect increases and decreases, within brainspace, in the ratio of oxygenated to deoxygenated blood. And this in turn enables strong inferences about where and when, in the brain, neurons are working harder.

2) *fMRI in Our Experiment, Specifically*

Let's recap, and apply to our experiment. Our experiment used fMRI technology, as just described, to scan a subject continuously as that subject sees each scenario stimulus on a screen, processes what it means, makes decisions about whether or not to carry a case, and registers that decision behaviorally, by pressing one of two buttons with her fingers.

Neurons work harder when a person is seeing, processing information, deciding, and pressing a button than when the person is not engaged in these activities. That calls up (so to speak) more blood, to deliver more resources. In the same way that transitioning from a jog to a sprint has our muscles calling up more oxygen and energy from the blood, neurons that are working harder call up more oxygen and energy as well.

Our subjects in the scanner saw repeated variations of the same basic case-carrying scenario. For each variation, they made a decision, and then conveyed their decision, about whether or not to carry the case they would receive in each scenario. And each scenario varied the probability (from 20% to 100%) that the case would contain contraband, and varied the probability (from 0% to 80%) that a carried case would be searched.

Throughout the entirety of these decisions, the scanner recorded data from the entire brain about where, when, and how oxygenated and deoxygenated blood ratios were changing. Because we knew exactly what each subject was seeing when, and knew exactly when and what the decision output (i.e. carry or don't carry) was, we could correlate different patterns of brain activity with different probability combinations, with different decisions each subject reported.

Each subject was in the scanner for just under 40 minutes. Since we measured changing oxygenated blood levels in tens of thousands of brain locations during that period, there were literally millions of pieces of data collected about each subject. Our next step was to analyze the data, which we did by deploying a form of artificial intelligence known as a machine learning algorithm (also sometimes in this context called multi-voxel pattern analysis) to which we next turn.

3) The Machine Learning Algorithm

"Machine learning" describes a process by which a software program can "learn" the associations between various inputs, conditions, and outputs. In our case, the inputs are the brain data. The conditions are such things as the separate risks, within each given trial, of carrying contraband and being searched. And the outputs are the subjects' choices whether to carry or not, given the conditions.

There is much more to it than this, as you may imagine. But the core idea for our purposes here is that if you train the algorithm by showing it a variety of actual data from actual subjects, the software first attempts to find the most common patterns within that data set. And it can then use those patterns to predict how to classify the subject's mental state, during a given trial, into a knowing or reckless frame of mind, using brain data alone. The machine "learns" what patterns of brain activity are associated with being in a knowing mental state by comparing the fMRI data gathered when subjects were contemplating carrying a single briefcase. And the machine "learns" how those patterns differ from the brain activity associated with being in a reckless mental state by comparing them with the brain activity when the subjects were contemplating a pool of 2, 3, 4 or 5 briefcases.

In fact, we used a particularly sophisticated algorithm, known as "the elastic net," that learned not just from the data, but also from the failures and successes of other efforts to learn from the data. We can clarify what that means with an analogy.

Imagine a teacher who, in her first year in the classroom, tries to teach a bunch of students to identify birds by showing them pictures. She puts a slide up on the screen and says, "Robin!" and then another and says, "Cardinal!" and then moves on to other slides of other species. Before her second year of teaching, she reviews the students' performance from the first year and finds that some of the pictures she showed were more useful than others for teaching the students. The students were confused by some pictures and found others more helpful. Perhaps they did exceptionally well at identifying cardinals shown from the front, on their final exam, and there is only one picture in the stack of a cardinal from that angle. From that result, she concludes that that one picture in the stack was particularly pedagogically useful. She repeats the process for other species, and makes extra copies of the useful pictures and adds them to stack.

Our hypothetical teacher then tries again the following year with a new group of students. They see all the original pictures, shown to the prior set of students, but the pictures that were useful last year they see more than once. Again the teacher reviews. She finds that even among those picture she made extra copies of, some were exceptionally helpful to the students. She makes yet further extra copies of those and adds them to the stack, creating a new, even better stack to use

for next year's students. And so on. In her tenth year of teaching, she has a great stack of photos, far better than her first-year stack. The tenth-year students, as a result, are fantastic at identifying birds from pictures, much better than the first-year cohort.

Our algorithm learned in a way analogous to this and so became better and better at classifying knowing and reckless mental states across several generations. Algorithms that work this way are sometimes called "pattern classifiers." And using such classifiers with respect to brain data is sometimes called "multi-voxel pattern analysis" or MVPA for short (where a "voxel" is like a 3-dimensional pixel volume in the brain, 2x2x2 millimeter cube).

Let's give a different example to make clearer how this can work, in a simple case. Suppose we wanted to see if a machine learning algorithm could reliably determine whether a person whose brain was scanned with fMRI was looking, at the time the brain data in question were acquired, at a photo of a face, or a photo of a place.

We could feed the algorithm brain data from when a bunch of different subjects are seeing faces, and "tell" the algorithm, essentially: "These data are all from condition 1, which we will call 'faces.'" We could then feed the algorithm brain data from a bunch of subjects who were at the time seeing places and "tell" the algorithm "These data are all from condition 2, which we will call 'places.'" Then we could show the algorithm new unlabeled brain data from a single subject and ask it to determine, on the basis of differences it observes between the two conditions, whether this person was in fact looking at a face or a place at the time the brain data were acquired.

The greater the differences between the aggregate sets of condition 1 and condition 2 brain data, the better will be the algorithm's ability to predict what the unknown subject was looking at. And in laboratory conditions, when researchers actually know what this mystery subject was looking at, but are testing the effectiveness of the algorithm, the accuracy of that prediction can be quantified (such as, say, 89% accurate). The more accurate the algorithm, the more confidence researchers can have about the predictions the algorithm can make with respect to subjects whose stimuli are *not* known to researchers. Consequently, if researchers are using a training method like the elastic net, they can then use their degree of confidence to alter their training method, emphasizing the particularly useful, and representative parts of the first round training data to retrain in the second round, in order to improve predictive power. And so on.

In like fashion, we set our algorithm the task of predicting whether one of our research subjects was in a knowing or reckless mental state, during any particular trial in the scanner. We also set our algorithm the task of predicting whether a subject in a reckless mental state was seeing 2, 3, 4 or 5 cases. And we also set our algorithm the task of predicting how many guarded tunnels

(representing search risk) the subject was seeing at a given moment. And we set the algorithm to predict whether or not, given the brain data observed, a subject was about to choose to carry the case, or decline to carry the case.

4) Testing the Machine Learning Algorithm

There are a variety of statistical techniques that can test the accuracy and reliability of a machine learning algorithm. We first used a common technique rather descriptively (if clunkily) called "leave-one-subject-out cross-validation."

There are more subtleties and complexities to this technique than we expect readers here will want to know.[17] But the key idea is that you can train the algorithm repeatedly, and independently, on one subset of the data you've already collected, and ask it to make predictions about the other subset. By continuously and precisely changing the subsets, you can get a very clear sense of the algorithm's accuracy.

For instance, if you have collected brain data on 40 subjects, you can have the algorithm learn from subjects 1 through 39, and then make a prediction about subject 40. Then you can start over, having the algorithm learn from subjects 2 through 40, and then make a prediction about subject 1. And so on and so on, always leaving one subject out, systematically varying which subject that is. This method of repeated testing gives clear indications of the algorithm's accuracy. If the algorithm does well in classifying the subject who was left out of the training set, no matter which subject that is, then that gives you greater confidence that the algorithm is tracking what it should be tracking.[18]

---

[17] Interested readers can find much more information on our methods for training the classifiers in *Predicting the Knowledge-Recklessness Distinction in the Human Brain*, *supra* note xx, and in the *Supplemental Information* published at the end of the article. Details on this particular method appear on page 4-5 of the *Supplemental Information*. Readers interested in learning more about such classifiers generally can see Kenneth A. Norman, Sean M. Polyn, Greg J. Detre, & James V. Haxby, *Beyond Mind-Reading: Multi-Voxel Pattern Analysis of fMRI Data*, 10 TRENDS IN COGNITIVE SCI. 424 (2006); Frank Tong & Michael S. Pratte, *Decoding Patterns of Human Brain Activity*, 63 ANN. REV. OF PSYCHOL. 483 (2012); Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, & Jack L. Gallant, *Encoding and Decoding in fMRI*, 56 NEUROIMAGE 400 (2011); John-Dylan Haynes, *A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives*, 87 PRIMER 257 (2015).

[18] We did some further work to assess the algorithm's accuracy, an appreciation of which requires that we introduce here, for more technically inclined readers, some additional subtleties about the way these algorithms work. So far, we have been speaking as though the post-training algorithm tells you, full stop, whether the brain data that you offer it was recorded from a reckless or a knowing subject. But, in fact, that's not what these algorithms produce. Rather, they provide you with *a degree of confidence* that the subject was reckless or knowing. They say that, for instance, there's a probability of .2 that the subject was knowing, or a probability of .75. They assign a number between 0 and 1 that represents the likelihood that the subject was in the same mental state as those in the training set, given what it learned from studying the training set.

E. Primary Findings

So, to recap, we asked our subjects to play a game while we scanned their brains. They each made 125 decisions as to whether to carry a briefcase across the border when given varying information about the probability that the briefcase contained contraband (the Contraband Risk) and the probability that the case would be searched at the border (the Search Risk).

We then built an algorithm—a digital machine, essentially—that takes brain data as an input and spits out one of two verdicts as an output: reckless or knowing with respect to the contents of the briefcase. The output is the machine's best guess about the mental state of the person whose brain data it takes as an input. We then used a variety of tools for measuring how well the machine worked. And we measured, that is, how well, using only information about a person's brain, it did the very job that we ask fact-finders to do whenever we ask them, using evidence admitted in court, to determine whether a criminal defendant was reckless or knowing.

What this means is that a further decision needs to be made in order to use the algorithm to actually classify subjects into the knowing or the reckless category: we need to decide how confident the algorithm needs to be in its classification before we will put the subject in the knowing or reckless category that the algorithm recommends. Do we want to classify the subject as knowing when the algorithm's confidence is above .3? How about .5? Or above .75? Or above .95? Or what? What's the appropriate threshold above which we pull the trigger and classify the subject as knowing (or reckless)?

Note that where-ever you place the threshold there will be inaccuracy that could have been avoided by placing the threshold elsewhere. If you place the threshold at .75, for instance, then subjects that the algorithm identifies as .6 will not be classified as knowing, even though quite a few of them were looking at a single briefcase when the relevant brain data was recorded. However, if you lower the threshold to .6, in order to classify them correctly, you will thereby misclassify those subjects who were merely reckless and who the algorithm assigned values between .6 and .75. Where-ever you set the threshold there will be false positives (reckless people who are classified as knowing), and false negatives (knowing people who are classified as reckless).

One question is where the optimal threshold is. At what threshold do you get the best mix of false positives and false negatives? This is a statistically soluble problem. Another important question, however, is how many choices of threshold provide you with a powerful classificatory tool? Does the algorithm do quite well when the threshold is set anywhere between .5 and .9, for instance? Or does it only perform well between .75 and .78? As a general rule, an algorithm used for classification is better if it is more robust, if it performs well for a wider range of choices of threshold. So that, itself, provides a measure of an algorithm's value. If it performs well over a wide range of choices of threshold, then that's a good reason to think that it is learning the right things from the training data. This, also, was part of our analysis. We assessed the value of the algorithm by seeing how robustly it provided accurate results over a range of thresholds.

Because we measured many different things, there were many different results. Here are the two most important ones, for immediate purposes.

1) Knowing and Reckless Brain States Differ

Our paramount finding is this: our algorithm correctly classified people as knowing or reckless 71% of the time, in some conditions.

Recall that prior empirical work has demonstrated that ordinary people, asked to classify people as knowing or reckless, are only slightly more likely to get a correct answer than we would get if we were to flip a coin. (That is, just above 50%.)

Our algorithm, by contrast, outperformed ordinary people, not to mention coin-flips, by a significant margin. And unlike ordinary people, who draw on a wide range of evidence about human behavior in making their decisions about another's mental state, the algorithm used only information about brain activity supplied by an fMRI.

It is this result that makes this experiment worth reporting to a legally-minded audience. In a sense, our algorithm was able (again, in some conditions) to read minds by looking at brains. And it did not read a trivial aspect of mind. For instance, it did not just distinguish between seeing a face and seeing a place. It read an aspect of mind crucial to mens rea, and so to criminal punishment.

Put simply: by combining fMRI brain-imaging techniques with a machine learning algorithm we were able to distinguish among guilty minds.

2) Order of Information Matters

Last section we twice indicated that we could make distinctions, on brain data alone "in some conditions." That is an important caveat, and one we wish to clarify immediately. And the caveat concerns the order in which subjects received risk-relevant information.

Recall that half our subjects were *first* presented with information about the size of the pool of briefcases from which their briefcase would be chosen and *then* next shown information about the likelihood that they would be searched at the border (we called this "the Contraband-First Condition"). The other half of our subjects saw these two pieces of information in reverse order (we called this "the Search-First Condition").

Interestingly, our algorithm was excellent at classifying the mental states of those in the Search-First Condition and abysmal at classifying the mental states of those in the Contraband-First Condition. Where, as just mentioned, the algorithm correctly classified subjects 71% of the time if they first saw the information about the likelihood of being searched, the algorithm succeeded in correctly classifying

only 32.1% of the time when examining information about the brains of those who first saw information about the likelihood that their briefcase contained contraband.

The difference between the two sequences in which subjects received information was also reflected in the behavior of our subjects. They were far less likely to choose to carry the briefcase across the border if they were presented first with the likelihood of being searched, and second with the likelihood that the case contained contraband, than if they saw the pieces of information in reverse order. This was true, importantly, when the probabilities of the various possible payoffs from carrying were held constant. Put another way: tell someone that they have a high chance of being searched, but almost no chance their briefcase contains contraband, and they are much less likely to choose to carry it than if you tell them that there is almost no chance the case contains contraband, but there is a high probability of being searched.

Part II. Implications of Detecting *Mens Rea* in the Brain

A. Immediate Legal Implications

The primary finding of our study has several important implications. First, our team's experiment provides a clear answer to the question: Does the distinction between knowing and reckless *mens rea* reflect a detectable distinction between brain states? The answer is: Yes.

On the basis of current evidence, the distinction is not simply projected onto people who are, considered in themselves, no different from one another. Put another way, the supposed distinction is no more in the eye of the beholder than detectable differences in the brain are in the eye of the beholder.

The alternative hypothesis, recall, is that the legal definitions of knowing and reckless do not apply differentially thanks to different psychological features of defendants. They instead reflect, on that view, independently formulated judgements about which defendants should be punished more severely. But that hypothesis is not consistent with the data we collected.

Using a combination of fMRI brain imaging and an algorithmic artificial intelligence, we were able to quite reliably predict – on the basis of brain activity alone -- whether or not a subject was in a knowing or reckless mental state. This suggests that differential liability can legitimately rest, if we retain our collective decision to do so, on there being a distinction between knowing and reckless mental states of the kind that is reflected in distinct neural activity. And that's because our main finding is, among other things, inconsistent with any argument that these distinctions are arbitrary, invented, or merely providing cover for juries or judges to punish some defendants more than others.

Although our main finding was not true in *all* conditions (recall that when subjects receive Contraband Risk information before receiving Search Risk

information the algorithm could not accurately distinguish the reckless from the knowing) the fact that it was true in *any* conditions strongly suggests (subject, of course, to future studies that may replicate and extend our findings) that there is a brain difference between those who are classified as knowing, on one hand, and reckless, on the other. It is possible, of course, that there is no meaningful difference between the knowing and the reckless in some conditions, and that there is a difference in others. But a more parsimonious hypothesis, given that we found an intrinsic difference in some conditions, is that there is a difference in others that fMRI cannot detect.

The Model Penal Code's assumption that those whom it classifies differently on the basis of their mental states actually differ psychologically has never before been directly tested. While we do not suggest that the results of a single study in any domain could ever lay a question to rest forever, our study should be seen as significantly increasing the likelihood that there are brain-based differences between people who are in knowing and reckless mental states. So the main implication of our study is that, whatever the relative merits of having or eliminating the distinction, calls for reform to eliminate the distinction are on considerably weaker ground, empirically, than they were previously.

Second, our results lend support for the idea that jurors need more help figuring out how to distinguish knowing from reckless mental states in real cases. As mentioned earlier, behavioral experiments in a separate set of published studies strongly suggest that jurors are quite poor at distinguishing these two mental states in the way the MPC insists they do. If our brain-imaging results had found no differences between the two mental states, and people can't reliably distinguish them anyway, then a concern for justice would recommend possible elimination of the distinction between the two. But if instead there are distinctions in the brain, and jurors have a hard time sorting defendants between the two, this recommends that we find a way to do a better job at instructing jurors how to sort accurately. If we are going to keep a system that punishes people in the knowing category more than people in the reckless category, then we had better ensure that jurors perform very significantly above chance when assigning defendants to one category or the other.

Third, our neuroscientific methods suggest the MPC mental state categories may not be nearly as unitary as currently supposed. That is, there may be important subcategories, and multiple subtypes, of culpable mental states. More specifically, our study suggests that the distinction between knowing and reckless mental states may be greatest when subjects perceive information about the presence or absence of an element of a crime after they come to have information about the likelihood of being caught; the distinction may be less obvious, or absent, when subjects perceive risks in the reverse order.

This suggests (though of course it does not yet prove) that keeping the knowing versus reckless distinction is far more salient when a subject first learns of the risks of getting caught doing a prohibited act than it is when a subject first considers the probability that his act will be prohibited. Or, put another way, we are on shakier ground, perhaps, in sorting some defendants into reckless and knowing than others. Those who are mistaken, for instance, about the illegality of their conduct—they think what they are doing is legal and so not subject to punishment—at the time that they commit a crime are possibly less good targets for classification into the categories of knowing and reckless than are those who know when they act that they are engaging in illegal activity, and so are at risk of being caught. There may be a far less meaningful distinction between knowing and reckless conduct when the actor is uncertain, or unaware, of the illegality of his conduct.

This, in turn, logically raises the question whether policy-makers should therefore consider keeping the knowing versus reckless bifurcation for some defined circumstances, and eliminating it for others. (We are not advocating this, or any other, legal reform, but rather are pointing out the possibility of such reform as a potential application of our finding.)

Fourth, our team's experiment and findings provide a concrete example of how neuroscientific methods can open up new avenues for discovering answers to some of law's enduring questions. On the one hand, we hasten to add that we are not zealots at the altar of a brain-scanning machine. We do not think that brain-scanning will entirely upend long-standing legal approaches to issues in either criminal law or civil law. We are pragmatists, observing the potential utility of new technology and associated methods. On the other hand, we believe this study clearly and amply demonstrates that there are some questions relevant to law as to which brain scanning can provide valuable new information. And the significance of this – entirely independent of the experiment's value in the substantive context of *mens rea* – should not be underestimated.

B. Implications for Future Work

The implications of our study extend beyond the boundary between knowing and reckless mental states. Our study points the way towards various future studies and avenues of research, each with legal implications of their own.

First, the line between knowing and reckless, which our experiment investigated, is only one mens rea line drawn by the Model Penal Code. And the Model Penal Code's divisions are not exhaustive of the mens rea distinctions drawn in American law. Similar studies, therefore, could be done to determine whether purpose and knowledge can be distinguished on the basis of brain data alone, or whether recklessness and negligence can be.

The Model Penal Code, also, for instance equates awareness of "high" probability with knowledge under certain circumstances (such as those in which the defendant does not "believe the [material element] does not exist").[19] Do we lose the ability to distinguish between those two under the special circumstances in which the Model Penal Code equates them, or not? And when we move beyond the Model Penal Code's mens rea regime we find various other questions that could be explored using the sort of tools we developed for this study. For instance, many jurisdictions in the United States reserve the most severe penalties for murders that are "willful, deliberate and premeditated". Is it possible to distinguish acts performed with that frame of mind, from those that are not, solely on the basis of brain data?

Second, our study specifically concerned knowledge and recklessness with respect to a circumstantial element of a crime—the presence or absence of contraband in the case, a fact that accompanies, but need not be caused by, the act of crossing the border. It is possible that we do not find the same, or any, brain-based difference even when it comes to other circumstantial elements of crimes.

Perhaps, for instance, the line between knowledge and recklessness when it comes to another's consent—the absence of which can also be a circumstantial element of a crime—cannot be drawn neurally, or must be drawn differently. Further work, that is, could investigate different forms of potentially illegal behavior also involving circumstantial elements. But, in addition, further work could expand beyond circumstantial elements to result and act elements of crimes. We do not know whether our results would extend to mens rea at the time of the act with respect to future harms that the act might cause.

Third, with further development our team's work could be extended to investigate the interaction of mental illness with criminally culpable mental states, about which we have almost no evidence-based knowledge. Except in those rare states that bar the use of evidence of mental disorder to negate mens rea, defendants routinely introduce evidence of the existence of certain recognized mental disorders –schizophrenia, post-traumatic stress disorder, autism spectrum disorder, and depression, for example – to raise reasonable doubt about the presence of some mens rea element of the crime. But there are to date no studies that directly examine the impact of mental disorders on mens rea.

Fact-finders receive some guidance from clinicians and forensic psychiatrists. But these experts' judgments are not supported by systematic, experimental findings. And it is not hard to see why. To investigate the question, for instance, of whether PTSD sufferers are more or less likely than the rest of us to know, as opposed to being reckless, about features of their environment that bear on their criminality, we would need a way of measuring, in lab conditions, which mental

[19] MPC 2.02(7).

state they are in, in comparison to healthy controls. Our study shows that tools for making such measurements can be developed, from combining existing fMRI technology with methods of artificial intelligence.

For similar reasons, our study shows that these tools can help to measure the impact of intoxicants on mens rea. Although there are significant limitations on how voluntary intoxication can be used to negate mens rea, most states allow defendants to shield themselves from liability on the grounds that due to intoxication they failed to know something, even if they would have known it had they been sober. There are many different intoxicants, of course, and they vary enormously in their psychological effects. Yet there is no data-driven work, akin to the experiment we've just described, that investigates the differential impact of, for instance, alcohol, cocaine, methamphetamine or marijuana on the "knowing" mental state. There now could be.

C. Non-implications

There are a variety of things that our experiment could be taken to imply that it does not imply. And these are important to highlight.

First, and most importantly, scientific findings never provide automatic support for a change in policy (or, conversely, a continuation of existing policy). So our findings don't either. Sound policy-making or policy-reform always requires that policy-makers view facts through the prism of values and consider them in light of fundamental normative principles. Put another way, there is no *automatic* pathway from description to prescription, or from explanation to justification. Facts warrant attention, of course. But whether or not they should inspire change depends on what it is that society is trying to accomplish, and what principles it must comply with in the effort.

In context, that means that if a state's statutory regime establishing different criminally culpable mental states is structured by and grounded on the assumption that there really are brain-based differences in those mental states, then facts supporting that assumption tend to increase our confidence in the regime. And facts inconsistent with that assumption tend to weaken it.

That said, legal regimes reflect a variety of values, and a variety of principled commitments. It is rare that we can honor our principles and at the same time maximize our values. Trade-offs must always be struck. And empirical results don't dictate how.

For instance, it requires payment from us all to produce and maintain the machinery of justice – which requires police, judges, lawyers, juries, prisons, and the like – much less to do so in a way that is faithful to fundamental principles of fairness. And perfect justice for every citizen would be prohibitively expensive, while reasonable expenses can supply only imperfect justice. We acknowledge (as

must all, we think, but it's useful to be explicit on this point) that finding the right or best balance is fiendishly difficult.

Likewise in the mens rea domain. Neither the results of our experiment, nor the results of any experiment, can alone answer the question whether we should or should not keep four categories of mental state, much less the four particular categories defined in the MPC. The fact that our experiment has found a brain-based distinction between knowing and reckless mental states cannot automatically justify the continued division of those states in the law, any more than would the absence of such a finding demand the elimination of the distinction. To be clear, the implication of our finding is not that the law must retain the knowing-reckless distinction; it is, instead, that *to the extent that the best policies require that the mens rea categories reflect differences in brain states*, our finding provides some support for maintaining the distinction. But whether the best policies require that is a profoundly difficult question that cannot be answered by doing experiments.

How many mens rea distinctions does justice require? And of what kind, dictating what differences in state treatment? Those are not questions that empirical investigations can answer. And without answers to them there is no saying how, if at all, mens rea law should be reformed. The facts about psychology, and about the neural substrates of our psychological states, can inform policy in this domain, but they cannot do the crucial work alone.

The second non-implication of our experiment is this: Our team's neuroscientific techniques can discover brain-based differences between mental states that exist at the time of scanning, *not at some prior time.* Although we have developed and deployed a powerful tool for exploring whether or not such differences exist, it is not (at least not so far) a tool for reliably exploring what mental state a subject was in minutes, hours, days, or even years beforehand. Put another way, our current experiment has implications for criminal justice policy, but not for forensic evaluation of individual defendants.

Third, the extent to which our study read the minds of subjects should not be exaggerated. True, it is remarkable, frankly, that the algorithm could classify subjects as knowing or reckless taking only information about their brains, collected non-invasively, into account. Such a thing would have been inconceivable 25 years ago. But that does not imply we now have a general-purpose mind-reading capability. No existing technology can yet (if ever) transcribe all the words or concepts a person is thinking. And readers should not mistakenly think that our technology simply peered into a subject's brain and could tell whether that subject was thinking in a knowing or reckless way.

Instead, our experiment showed that there were sufficiently great differences between knowing and reckless brain activity that the combination of fMRI and artificial intelligence could *learn* that difference (not just – on its own – *discover* and *name* the difference). The crucial distinction, and the point we are

emphasizing here, is that human instructors had to provide the algorithm with two *potentially* different conditions to examine, in the first place. Specifically, we instructed the algorithm to look for differences between two experimental conditions (i.e., the one-case condition, on one hand, and the more-than-one-case condition, on the other) that we believe separately invoke knowing and reckless mental states. Had we not asked the algorithm to look for differences in subjects between these two conditions, it would not, on its own, have looked for (or thereby found) any. Just as a student could never learn to identify a robin from a picture unless the instructor already knew how to identify robins, no algorithm can learn to identify a knowing actor from her brain activity unless its "teacher" can independently identify the knowing actors. Typically, and maybe even always, machines cannot learn to do things that people cannot already do (perhaps more slowly or less efficiently) without their help. That is just as true of machines taught to "read minds" from brain activity. Clarifying that should dispel any mistaken conclusion that mind-reading of the expansive sort is now available.

Fourth, it is worth noting that we have not yet said in this article what kinds of activities, in which particular brain structures, enabled our algorithm to distinguish between subjects in knowing and reckless states of mind. The reason we have said nothing about this so far is not that our study has nothing to say. It does.[20] The region known as the dorsolateral prefrontal cortex (dlPFC), for instance (a region known to be involved in planning, analysis, and deliberation) was among the regions of the brain that behaved distinctively in subjects in states of knowledge. And, as mentioned above, there may be times – such as when investigating the impacts of mental illness or intoxication on legally relevant mental states and behaviors -- when knowing which regions are actively involved in which legally relevant behaviors may be quite fruitful to know.

However, for purposes of our specific research question – whether knowing and reckless mental states are distinguishable in the brain – the locations of differences is simply less legally relevant than the fact that discernable differences exist. That is, the central result that we reached—distinguishing knowledge from recklessness solely on the basis of brain data—is significant quite independently of what aspects of the brain made the result possible.

[20] Interested readers can find details in our *Predicting the Knowledge-Recklessness Distinction in the Human Brain*, *supra* n. xx. In brief, areas more predictive of being in a knowing situation included the anterior insula (often involved in risk and uncertainty representation), dorsolateral prefrontal cortex (associated with executive decisions and computation) and temporo-parietal junction (often involved in moral decisions). Areas more predictive of being in a reckless mental state include the occipital cortex (sometimes involved in circumstances of high uncertainty). *Id*. at 3-4, and *Supplemental Information.*

Part III. Cautions & Caveats

The brain imaging method we used – fMRI – is a fairly recent technological advance, and a remarkable technique for learning about brain activity in a relatively non-invasive way. For this reason, publication of MRI and fMRI studies from major universities (which can pay several million dollars for a high field-strength machine) has exploded. For instance, a literature search in widely-used PubMed database revealed that although in 1987 less than 200 articles using these methods were published each month by 2014 that figure was typically above 2000 such articles per month.[21] A 10-fold increase in 26 years. Looking just at fMRI publications, a 2010 study in the same database found a rise in annual publications from effectively 0 in 1992 to well over 2000 annually in 2009.[22] And the excitement within the neuroscience community, over the prospects fMRI offers for continuing discoveries, is quite palpable.

At the same time, we want readers to understand that we pitch down the middle – neither more zealous nor more skeptical than fMRI is due. Studies by our working group, by other working groups in our Network, and by other research teams around the world, have demonstrated that neuroscientific techniques can add value to law's efforts. But brain-scanning is not magic. It has limitations, many of which we have helped to explore and detail.[23] For this reason, we believe it is appropriate to lay on the table, transparently, a variety of cautions that might help readers to strike the right balance between under-interpreting and over-interpreting the specific findings we describe here, as well as fMRI studies in general.

---

[21] Nikki Marinsek, *20 Years of trends in the MRI and fMRI Literatures*, at https://nikkimarinsek.com/blog/fmri-bursts (last visited July 14, 2019).

[22] Lars Muckli, *What Are We Missing Here? Brain Imaging Evidence for Higher Cognitive Functions in Primary Visual Cortex V1*, 20 INT'L J. OF IMAGING SYSTEMS & TECH. 131, 132 (2010). Independent analysis by authors, confirming this, on file with authors.

[23] *See, e.g.*, Jones, Owen D., Richard J. Bonnie, B. J. Casey, Andre Davis, David L. Faigman, Morris Hoffman, Read Montague, Stephen J. Morse, Marcus E. Raichle, Jennifer A. Richeson, Elizabeth Scott, Laurence Steinberg, Kim Taylor-Thompson, Anthony Wagner, and Gideon Yaffe, *Law and Neuroscience: Recommendations Submitted to the President's Bioethics Commission* , reprinted in 1(2) JOURNAL OF LAW AND THE BIOSCIENCES 224 (2014). *See also* Chapter 9 – *Limits and Cautions*, in Owen D. Jones, Jeffrey D. Schall, Francis X. Shen, LAW AND NEUROSCIENCE (2014) and Owen D. Jones, Joshua Buckholtz, Jeffrey D. Schall, & Rene Marois, *Brain Imaging for Legal Thinkers: A Guide for the Perplexed*, 2009 STANFORD TECH. L. REV. 5; Russell Poldrack, *The Role of fMRI in Cognitive Neuroscience: Where do we Stand?*, 18 CURRENT OPINION IN NEUROBIOLOGY 223 (2008); John T. Cacioppo et al., *Just Because You're Imaging the Brain Doesn't Mean You Can Stop Using Your Head: A Primer and Set of First Principles*, 85 J. PERS. SOC. PSYCHOL. 650 (2003).

Before describing a few cautions specific to fMRI, however, it is worth reminding readers of several kinds of cautions that apply to any laboratory experiment, and therefore to ours as well.

A. General Cautions

First, there is always a trade-off between how realistic (or "ecologically valid") an experiment is and how well the circumstances different subjects encounter can be controlled. The real world is more authentic, of course. But it is also so messy in terms of constantly shifting variables – such as temperature, body language, ambient sound and lighting, verbal cues, stimulus duration, and the like – that conclusions are weakened by potentially unknowable susceptibilities to uncontrolled variables that may in fact be hidden causes of anything interesting in a study's results. Conversely, a great many variables can be controlled (that is, held constant) in strict laboratory environments. But that very control could narrow the generalizability of a study's findings, which might only be true in identically controlled environments.

Given that there is as yet no reliable way to record brain function as a person is going about her day, and perhaps encountering situations that might (or might not) invoke law-relevant mental states, our experiment obviously had to be conducted in a laboratory. And although brain scientists have no reason yet to believe that brains operate very differently outside the lab than inside, transparency requires that we at least mention the possibility. The choices our subjects made in the scanner about whether or not to carry a case, given particular monetary payoffs, may not be exactly the same as the choices they would make in the real world, with real cases laid before them, less certain probabilities of outcomes, and large sums of money to be gained or lost. Fortunately, however, our concern in this experiment is not with most accurately identifying what a criminal's brain activity is like, during the moment of criminal activity, but rather with the narrower question of whether there are any detectable differences – in the scanner at least – between knowing and reckless decision-making, when all else is held constant.

Second, although our sample size of 40 subjects is within the norm, in fMRI brain imaging studies, for investigating brain activity with sufficient statistical power to publish findings in top peer-reviewed neuroscience journals, there is always the possibility that a larger study would find either fewer or more differences between knowing and reckless mental states.

Third, there is always a possibility that sampling different demographic groups will yield different results. For instance, and as is typical of brain-scanning experiments, our subjects all came from the geographic region near the scanning facility – in this case the Roanoke/Blacksburg Virginia area. Would the brain activities of senior citizens in North Dakota, gamblers in Las Vegas, or police

officers in Los Angeles be different? To what extent does education, sex, nationality, nutrition, health, or socioeconomic status affect the results? There is at present no reason to believe that different groups will use their brains quite differently; but we would be remiss not to mention the possibility.

B. fMRI-Specific Cautions

There are several key limitations to fMRI techniques, of which readers should be aware.[24] First, fMRI is an indirect, rather than direct, measure of neuronal activity. Instead of measuring the electrical activity of individual neurons, or even a group of them, fMRI detects changes in blood oxygenation levels, over time, in discrete locations within a subject's brain that include neurons, as well as other brain tissue. There is every physiological reason to believe that the more various neurons fire, the more resources (such as oxygen and glucose) they demand. Still, it is a little like distinguishing cities from countryside by measuring differential regional light outputs from space at night. In the same way that that would measure something very reliably associated with cities, but would not be a direct observation of cities themselves, fMRI measures something very reliably associated with neuronal activity, without measuring the neuronal firings themselves.

Second, fMRI cannot identify differences between the kinds of neurons that are active. fMRI compares total activity within voxels (which are, as mentioned earlier, a cubic volume of brain tissue). But each voxel contains a great many neurons in number—usually estimated as over 600,000—and can also contain many different types of neurons. Some neurons, for instance, fire in a way that activates many other neurons in turn. But some kinds of neurons fire in a way that inhibits the activation of other neurons. Because fMRI does not distinguish among these sometimes competing purposes of neurons, it is akin to recording the decibel level in a crowd of people, many of whom are yelling "go", some of whom are yelling "no", and some of whom are keeping quiet.

Third, the fMRI brain images that researchers (including our team) present and publish are not at all like familiar x-ray images, which are the direct result of imaging technology interacting with brain tissue. Each image is, instead, something called a statistical parametric map. That is a combination of a structural image (akin to an x-ray image) of a single, typical brain onto which has been overlaid a patchwork variety of colors, in various locations, that represent the voxels with the most statistically significant differences between conditions. (Such as, in our case, the most significant differences in activity between knowing and reckless conditions.) The colors are calibrated to the range of greater and lesser differences. (The data conveyed in this fashion can be, and often is, also presented

[24] *See generally Brain Imaging for Legal Thinkers, supra* note xx.

in a table of figures, displaying coordinates of the particular brain regions that have the most significant differences in activity between tasks.) The key point is that the pretty pictures people see of colorful brain images is not a direct measure of where the brain is "lighting up." It is instead a color-coded representation, designed to be more quickly and intuitively grasped, of *where* in the brain there were differences between conditions, and *how great* those differences were.

Fourth, fMRI can (among other things) detect differences between various brain states, when subject brains are evaluating different scenarios and deciding what to do. Yet detecting brain state differences, on one hand, and discerning either the specific causal pathways for, or the meanings of, those differences is a different matter entirely. It is a very important step, but it is still a step on a journey, rather than the arrival at a final destination.

C. Algorithm-Specific Cautions

As tempting as it can be to laud the breakthrough capabilities of partnering machine-learning algorithms with brain-scanners, it is also important for legal thinkers to see their limitations with a clear eye, and to not succumb to temptation to overinterpret results. Specifically, we believe that although multivoxel pattern classification is a powerful tool for identifying the existence of salient brain differences, it will rarely provide strong support for claims about either: a) the precise *function* of any brain region; or b) any brain region's *centrality* to any particular and complex form of psychological functioning.

We can illustrate the source of our concern with an example. Say, for instance, that in Subject 1, regions A and B are active when he views a photograph of a sunset, while in Subject 2 it is regions B and C, and in Subject 3 it is regions A and C. None of these regions are active, let's stipulate, in subjects who are not seeing a sunset.

Suppose we now train an algorithm, with data from Subjects 1, 2, and 3 to identify when people are viewing sunsets. We then test the algorithm's ability by asking it to predict whether Subject 4, in whom only region A was active, was viewing a sunset at the time.

The algorithm will be quite confident that Subject 4 is viewing a sunset. After all, his brain is more like that of the sunset-viewers than it is like the brains of those who are not. Nevertheless, and here's the first key point: it would be false to conclude from this that region A is the "sunset viewing region of the brain." Second key point: it would be false to conclude that activity in region A is essential for viewing sunsets. After all, region A is not active when Subject 2 views a sunset; so activity there is not essential for viewing a sunset. Further, no other region is active in Subject 4, even though all the other sunset-viewers in our sample had other regions active when they viewed sunsets.

The key caution to keep in mind is this. Although algorithms are capable of learning to apply complex, disjunctive rules for classification, rules of that kind are not necessarily useful for gaining insight into the basic psycho-physical laws that govern the relationship between brain activity and psychological states.

D. Experiment-Specific Cautions

Against the background of the foregoing, there are several reasons, specific to our experiment in particular, to be cautious. First, the stimuli we used in the lab to elicit the mental states to be studied may or may not elicit those mental states perfectly. We strongly believe that the essence of the distinction between knowing and reckless mental states, as envisaged under the MPC, reflects different probabilities – such that: a) judging there to be a 100% likelihood that something will happen is "knowing" that it will; and b) lesser likelihoods reflect lessening degrees of "recklessness." Nevertheless, one could always argue that manipulating certainty and uncertainty in the laboratory misses something essential about the MPC mental states.

Second, context may matter. It is possible that the particular brain-based distinctions our team has identified between knowing and reckless decisions, when deciding whether or not to carry contraband, may not generalize to all knowing and reckless decisions, when deciding whether or not to engage in other kinds of activities. It is possible that there may be different distinctions, or even theoretically no distinctions, between knowing and reckless decision-making in other contexts, such as paying/avoiding taxes, or driving over/under the speed limit, or shooting/not-shooting a gun, and the like. Further work, as we indicated above, could investigate this issue. But in the absence of such further research, there is reason to be cautious.

Third, it seems appropriate to acknowledge that as researchers we can't be certain what a subject actually "knew" in the scanner. We told subjects that one (and only one) case each trial would contain "valuable content." And we told them that each trial they would be presented with between one and five cases. Therefore, subjects shown only one case on a given trial could straightforwardly deduce that that case logically *must* contain the valuable content, because the probability that it would do so is 100%. But that does not mean that we researchers knew that the subject knew the case contained valuable content. True, we did see behavioral differences in expected directions when only one case was on offer. Subjects, for instance, were certain to carry the case when there was only one offered and the probability of detection was zero. But we cannot claim that all subjects actually knew, on any or all such one-case trials, what they clearly should have known.

Conclusion

Pick any weekday, and you will find thousands of criminal trials underway in America. In almost all, the justice system tasks lay citizen jurors to decide if the accused did something prohibited, while in a culpable state of mind. In states that follow the Model Penal Code (MPC), which is the supermajority of states, there are four culpable states of mind: purposeful, knowing, reckless, and negligent. Each mental state bears a technical legal meaning, despite a common lay meaning that travels the language in parallel.

The law predicates differences in criminal liability on what the law supposes to be independently specifiable psychological differences that underlie and constitute differences in criminal culpability. But is this presupposition true? If there are such psychological differences, there must also be brain differences. Consequently, the moral legitimacy of the Model Penal Code's taxonomy of culpable mental states – which punishes those in defined mental states differently – depends on whether those mental states actually correspond to different brain states in the way the MPC categorization assumes.

The experiment described here is the first to investigate whether one long-standing assumption underlying the Model Penal Code's approach to culpable mental states stands up to empirical scrutiny. More specifically, we and our colleagues coupled fMRI brain imaging techniques and a machine learning algorithm (a form of artificial intelligence) to see if the brain activities during *knowing* and *reckless* states of mind can ever be reliably distinguished.

As our experiment indicates, the answer is Yes. Not only does our experiment provide a concrete example of how neuroscientific methods can contribute information relevant to legal policy. First and foremost, our experiment demonstrates that it is possible to predict, with high accuracy, which mental state a subject is in using brain imaging data alone. The results of our experiment therefore provide the first empirical support for the law to draw a line between, and to establish separate punishment amounts for, knowing and reckless criminality. This discovery could be the first step toward legally defined mental states that reflect actual and detectable psychological states, grounded in neural activity within the brain.

Our results do not by themselves suggest that there should or shouldn't be legal reform with respect to the presence of these two, or any other, mental states in a criminal law regime. Put another way, we have provided evidence that knowing and reckless mental states are – at least in some contexts – different in the brain. And we believe that that information is valuable if one cares about whether or not the MPC's approach to culpable mental states can bear the weight that the law asks it to. But our finding that the mental states can indeed reflect different brain activity does not mean the MPC distinction between knowing and reckless states should remain intact, any more than a contrary finding would mean

that the MPC distinction between knowing and reckless mental states should be abandoned. Support for policy change comes from the intersection of values, facts, and fundamental principles, and not from the facts alone.

We should all be interested in evidence-based legal reforms. But such reforms require evidence on which they can be based. When it comes to the law of mens rea, the relevant source of such evidence is psychology, cognitive science and, thanks to increasingly sophisticated technology for measuring the brain, neuroscience. As we hope to have demonstrated here, when neuroscientific techniques are aimed directly at questions of legal relevance, they can provide exactly the kind of evidence that can aid, although not dictate, intelligent, thoughtful legal reform.