Intersect, Vol 16, No 1 (2022)

'Neuralizing' Injustice: How neuroscience misunderstands racism, addiction, and crime

Chris Sanjeev Iyer
Stanford University

Abstract

The science of the brain has emerged to the forefront of public thought and policy regarding many social issues, enabled by technological advancements in cognitive neuroscience and psychology. While this research presents exciting possibilities for informing social activism and policy, centering neuroscientific explanations for complex social issues often obscures social injustices not easily measurable with neuroscientific tools. Here, I discuss neuroscientific and psychological research into racism, drug addiction, and criminality that has 1) *'over-individualized'* these issues by neglecting structural and environmental complexities, and 2) *'over-neuralized'* these issues by reducing them to neural phenomena, inscribing historical injustices into the 'hardwiring' of the brain under a veneer of scientific objectivity. Taken together, this research—on implicit racial bias, the 'brain disease' model of addiction, and the 'criminal brain'—casts the brain as a moral scapegoat, allowing us to show mercy towards individuals without grappling with collective responsibility for the conditions of injustice at the heart of racism, addiction, and crime. Finally, I discuss recommendations for conducting neuroscientific and psychological inquiry that is attentive to non-neural explanations and responsive to how the science of the brain is translated and communicated into broader society.

Introduction

In the 1990s—the so-called "Decade of the Brain"—much investment, interest, and hope was channeled into the rapidly proliferating field of neuroscience.[1] As the field's explanatory power boomed, some researchers began to explore *social neuroscience*—i.e., how neuroscientific methods might shed new light on social phenomena, rather than just biomedical ones.[2] In large part, this advancement was enabled by new neuroimaging

[1] Poeppel, D., Mangun, G. R., & Gazzaniga, M. S. (Eds.). (2020). *The Cognitive Neurosciences* (6th ed.). MIT Press.

[2] Cacioppo, J. T., & Berntson, G. G. (1992). Social psychological contributions to the

technologies that can measure activity in the *functioning human* brain (as opposed to post-mortem dissections or non-human animal models).[3] Researchers could now observe the neural structures and patterns of function that are implicated in socially consequential mental illnesses like PTSD,[4] in learning new educational information,[5] or in making different kinds of moral decisions,[6] to name a few.

In turn, neuroscience has revolutionized public thought and policy in diverse spheres of society: for example, how we treat mental illness and trauma, how we design classroom curricula, and how we conduct criminal trials. It offers us a *multilevel* understanding of such social phenomena, informed by the intricacies of cognition and the complex patterns of behavior that are written onto our brains.[2] The social promise and value of this multilevel understanding is undeniable.

Amidst this optimism, however, many overlook the ways in which myopically focusing on the brain has distracted us from social injustices not measurable with neuroscientific tools, and even in which neuroscience has been contorted into an instrument of dehumanization. Here, I will examine how neuroscience researchers have approached three pernicious social issues: racism, addiction, and criminality.

Two cross-cutting themes will guide my discussion of how research in these areas has entered into social discourse and policy. First, in seeking to operationalize social phenomena into constructs accessible with neuroscientific methods, researchers have often '*over-individualized'* these phenomena (and consequently, our proposed solutions to them). The tools of neuroscience and cognitive psychology are ill-suited to make claims beyond (groups of) individuals—that is, about social structures and institutions. Consequently, neuroscience's individual-level explanations can obscure structural and environmental aspects of racism, addiction, and criminality.

Second, applying neuroscience to these issues has *'over-neuralized'* our understanding of them, tempting us to believe that racism, addiction, and criminality should be understood (and therefore, intervened upon) as features of our brains—not just divorced from social conditions, but

---

decade of the brain. Doctrine of multilevel analysis. *The American Psychologist*, *47*(8), 1019–1028. https://doi.org/10.1037//0003-066x.47.8.1019

[3] Ochsner, K. N., & Lieberman, M. D. (2001). The emergence of social cognitive neuroscience. *The American Psychologist*, *56*(9), 717–734.

[4] Pitman, R. K., Rasmusson, A. M., Koenen, K. C., Shin, L. M., Orr, S. P., Gilbertson, M. W., Milad, M. R., & Liberzon, I. (2012). Biological studies of post-traumatic stress disorder. *Nature Reviews*

[5] Thomas, M. S. C., Ansari, D., & Knowland, V. C. P. (2019). Annual Research Review: Educational neuroscience: progress and prospects. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *60*(4), 477–492. https://doi.org/10.1111/jcpp.12973

[6] Greene, J. D. (2015). The cognitive neuroscience of moral judgment and decision making. In *The moral brain: A multidisciplinary perspective* (pp. 197–220). Boston Review. https://doi.org/10.7551/mitpress/9988.001.0001

divorced even from *ourselves* as empowered agents. This last portion is critical; in each of these three issues, a fallacious divide between the brain and the self has emerged, allowing us shift responsibility for social injustice from ourselves, as a collective, onto our brains. As I will discuss, this shift casts the brain as a moral scapegoat, allowing us to show mercy towards individuals without grappling with collective responsibility for the conditions of social injustice at the heart of racism, addiction, and criminality.

Throughout, I use the term *brain science* to refer more generally to cognitive psychology and neuroscience—avenues of inquiry broad enough to study the intersection of the brain with its surrounding social environment (i.e., not strictly neurobiology or neurochemistry), but specific enough to be still primarily concerned with brain functioning (i.e., not social psychology or sociology). While not all such researchers directly measure brain structure or activity, these fields are primarily concerned with the functioning of the brain, even if measured through behavioral, computational, or other assays. I refer to these fields jointly as *brain science* to indicate their continuity and confluence into 'neuro-informed' policy.

In part one, I describe brain science's main contribution to anti-racism efforts: implicit bias. I discuss how psychological and neuroscientific research on implicit bias falsely operationalizes racism as a trait of the individual—further, of the individual brain. This individual-level inquiry, in turn, distracts from *structural* racism and alleviates some collective responsibility for racial injustice. In part two, I take up the dominant neuroscientific view of drug addiction, the *brain disease* model, and how narratives emerging from brain science research obscure the social and environmental dimensions of drug use and abuse, instead spawning 'over-neuralized' misunderstandings of addiction that have contributed to excessive stigma, overdose epidemics, and mass criminalization. In part three, I discuss the concept of the *criminal brain*—how researchers have sought to operationalize (1) criminal behavior and risk, in a way that promotes dehumanizing punishments and reinscribes racialized notions of criminality onto the brain, and (2) criminal culpability, in a way that dissociates brain and self, once more distracting us from collective responsibility for the social conditions that (re)produce violence.

Across these topics, brain science research—and its public communication and social application—has centered individual- and brain-level understandings of phenomena which are far more complex than the tools of brain science can reach. In doing so, it has often inadvertently supported social injustice, described by sociologist Erik Olin Wright and political scientist Joel Rogers as "an inequality which is unfair and *which could be remedied* if our social institutions were different."[7] I discuss instances of social marginalization, stigmatization, and

[7] Wright, E. O., & Rogers, J. (2010). *American Society: How It Really Works*. W.W. Norton.

punishment that inflict social harm, do so unequally, and could be remedied in part by conscientious scientific inquiry and advocacy. As we will see, in advocating for care for drug-addicted individuals or mercy for criminal defendants, shifting blame from the individual to the brain gets us no closer to clarifying and intervening on injustice in underlying social conditions—interventions that brain science might play a crucial role in, if the field learns from the pitfalls discussed here.

I. Racism and Implicit Bias

Cognitive psychologists and neuroscientists have predominantly engaged with racism through the concept of *implicit bias*. In the 1990s and early 2000s, researchers began to develop psychometric tools to measure unconscious biases in psychological attitudes—perhaps most notably, the Implicit Association Test (IAT).[8,9,10] In the IAT, subtle differences in response times to various words or pictures in an associative matching task reveal implicit mental associations—for example, between 'black' and 'dangerous,' 'woman' and 'domestic,' or even just between 'flowers' and 'pleasant' (for more details, see Banaji, 2001).[8] These associations, as well as the neural connections that underlie them and the behaviors that arise from them, are described as *implicit biases*. The IAT has revealed various common biases, including ones that we might not consider harmful (like the latter above). However, the race IAT has received particular attention, especially in motivating social interventions like implicit bias training. Crucially, one's *implicit* attitudes (as measured by the IAT) can be biased even when someone professes *explicit* egalitarianism.[11]

The IAT contributed to a broader scientific movement to measure implicit biases with the tools of brain science. For example, one line of research studies the 'Own-Race Bias' (ORB) in memory, describing people's tendency to better remember faces perceived to be of their own race (similar research extends to 'Own-Gender Bias' and 'Own-Age

[8] Banaji, M. R. (2001). Implicit attitudes can be measured. In *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 117–150). American Psychological Association. https://doi.org/10.1037/10394-007

[9] Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6). https://doi.org/10.1037//0022-3514.74.6.1464

[10] Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295x.102.1.4

[11] Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The Rules of Implicit Evaluation by Race, Religion, and Age. *Psychological Science*, *25*(9), 1804–1815. https://doi.org/10.1177/0956797614543801

Bias').[12,13,14] Cognitive neuroscientists add to these psychological accounts, for example, by showing differential brain activity in face-responsive brain regions when viewing own- and other-race faces, as well as the triggering of fear-responsive brain regions while viewing other-race faces.[15,16,17] These neural biases correlate with psychological measurements like the IAT or ORB, offering potential neural mechanisms for implicit bias. Through implicit bias, the individual-focused tools of cognitive psychology and neuroscience—behavioral measurements like response times and memory accuracy, as well as neuroimaging of brain activity—became revelatory in conversations of racism and racial inequality.

As research on implicit bias proliferated, accounts of where these biases arise from became necessary. Here, researchers began to hook up individual-focused notions of implicit bias to broader social phenomena. For example, the 'perceptual expertise' hypothesis—one representative account for implicit racial bias in face processing—claims that because our communities and daily life are often organized along racial lines, we encounter more members of our own race; therefore, our visual systems become tuned to own-race faces.[12] Similarly, implicit associations measured by the IAT might arise from stereotypes and biases in our social environments—for example, from disproportionate media portrayals of black people as criminals (in fact, universal exposure to these stereotypes can explain implicit biases against one's own racialized group).[18] Our brains tune their processing to statistical regularities in our environments, so biases in our environments etch themselves into the way we think, perceive, and behave in our racialized social world.[19]

---

[12] Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law*, *7*(1), 36–97. https://doi.org/10.1037/1076-8971.7.1.36

[13] Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition*, *21*(9–10), 1306–1336.

[14] Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin*, *138*(1), 146–174. https://doi.org/10.1037/a0025750

[15] Golby, A. J., Gabrieli, J. D. E., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience*, *4*(8), 845–850. https://doi.org/10.1038/90565

[16] Kubota, J. T., Banaji, M. R., & Phelps, E. A. (2012). The neuroscience of race. *Nature Neuroscience*, *15*(7), 940–948. https://doi.org/10.1038/nn.3136

[17] Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation. *Journal of Cognitive Neuroscience*, *12*(5), 729–738. https://doi.org/10.1162/089892900562552

[18] Dixon, T. L., & Linz, D. (2000). Race and the Misrepresentation of Victimization on Local Television News. *Communication Research*, *27*(5), 547–573. https://doi.org/10.1177/009365000027005001

[19] Anderson, J. R., & Schooler, L. J. (1991). Reflections of the Environment in Memory.

From this research, a story of implicit bias emerged. Racial bias shapes, and can be measured in, our implicit attitudes, mental associations, and neural connections—the fundamental bases of how we think and act. We forge biased mental associations and manners of processing, and we behave in biased ways even when we profess (and often truly believe in) egalitarian attitudes. These biases emerge from difficult-to-uproot biases in our social environments, although often (especially in public discourse) their source is not considered at all. Crucially, with measures of explicit prejudice declining and failing to explain persistent accounts of institutional racial inequality, this story of implicit bias allows us to preserve racism as an individual trait—an individual attitude entirely unbeknownst to us, whose consequences reverberate throughout the social institutions we create.[20,21] However, this molding of racism around implicit bias, albeit convenient for explaining the disconnect between explicit prejudice and systemic racism, falls short—both in its scientific robustness and its social ramifications.

Scientifically, implicit bias has not held up to be a reliable and valid individual trait. We might expect that each person's implicit racial bias (as measured by the IAT) should be roughly stable over time; however, researchers have noted a low test-retest reliability of IAT scores, indicating that one's performance on the IAT changes drastically, even in the span of minutes.[21] Also, if measures like the IAT hold explanatory potential, then biases in response times measured by the IAT should be predictive of racially biased behaviors themselves. However, even the IAT's pioneers admit that at most 5.5% of discriminatory racial behavior (measured, for example, with person perception judgements or discriminatory social policy endorsements) is accounted for by race IAT scores, and critics of the IAT put this number below 1%.[22,23] Implicit bias—as measured by psychometric tools like the IAT—defies expectations we might have of an individual trait, and it fails to predict the racially biased behaviors and social practices that motivated its entry into public discourse about racism.

Psychological Science, *2*(6), 396–408. https://doi.org/10.1111/j.1467-9280.1991.tb00174.x

[20] Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*(4), 553–561. https://doi.org/10.1037/pspa0000016

[21] Payne, B. K., & Hannay, J. W. (2021). Implicit bias reflects systemic racism. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2021.08.001

[22] Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

[23]Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 171–192. https://doi.org/10.1037/a0032734

In response to this evidence discounting implicit bias as an individual-level trait, cognitive psychology researchers began to wonder if implicit bias in fact tells us more about bias in *environments and social structures* than it does about bias in individuals. One team found that while individual IAT scores are not stable over time, *state-level averages* are; and in fact, these averages correlate strongly with measures of systemic racism like racial disparities in police killings, or even Google searches of racial slurs.[21] They further point out that test-retest unreliability of IAT scores might reflect how the accessibility of mental associations, including those measured by the IAT, varies from context to context.[19,24] In highly discriminatory environments, harmful racial stereotypes might become more salient in our minds, driving up individual IAT scores. Perhaps, as these researchers argue, implicit bias has been wrongly described as an individual attitude, when in fact it is a cognitive expression of *systemic* racism beyond the individual level.

To summarize, neuroscientists and psychologists began to study *implicit bias*, a manifestation of racism that the individual-focused tools and perspectives of brain science can measure. However, inattention to social environments led these researchers to falsely construe implicit bias as a trait of the individual brain, when in reality it may better reflect racism in the same environments and structures that brain scientists often seek to 'control' for. This misconstrual has potent ramifications, both within and beyond science.

Scientifically, failing to attend to social conditions (and histories that produced them) perpetuates a misunderstanding of the brains and individuals under study. Brain scientists often seek to 'control out' the complexities of racial experience (e.g., unequal educational experiences or mental illness) to focus on an operationalization of racism (i.e., implicit bias) that falls within the individual-focused purview of brain science.[25] As Oliver Rollins writes, by controlling for aspects of racism that do not fit neatly within the box of implicit bias, "researchers risk reproducing scientific racism through the omission of racial experiences that do not fit or are too tricky to understand, in neurobiological calculations."[25] Thus, as researchers now challenge the view of implicit bias as an individual trait, it will be important to grapple with and clearly communicate the limitations of brain science to capture the socially and historically complexity of racism.

Beyond the scientific community, socially inattentive research contributes to public misunderstandings of racism as an individual-level phenomenon. A 2016 Pew Research Center survey found that more

---

[24]Dasgupta, N. (2013). Implicit Attitudes and Beliefs Adapt to Situations: A Decade of Research on the Malleability of Implicit Prejudice, Stereotypes, and the Self-Concept. *Advances in Experimental Social Psychology*, *47*, 233–279. https://doi.org/10.1016/B978-0-12-407236-7.00005-X

[25] Rollins, O. (2021b). Towards an antiracist (neuro)science. *Nature Human Behaviour*, *5*(5), 540–541. https://doi.org/10.1038/s41562-021-01075-y

Americans believe racism to be best understood as an individual attitude and interpersonal issue than a structural one,[26] despite a discrepancy between increasing egalitarian attitudes and worsening (or stagnating) measures of structural racism.[27] Centering individual-level accounts of racism in public policy and media, bolstered by implicit bias science, distracts from structural accounts.

Moreover, this over-individualized account of racism as defined by implicit bias supports 'bad apple' politics and risks further entrenching racial discrimination. Institutions like police departments, corporations, or courtrooms might deploy the 'science of racism' (i.e., implicit bias) in the form of implicit bias trainings, in order to argue for their long-term viability in the face of mounting criticism for racist outcomes.[28,29] Such trainings show little evidence of effectiveness (perhaps because of social inattention in the underlying research constructs) but receive much public attention.[30,31] As Khalil Muhammad notes, "implicit bias is not the whole problem, nor does it alone change the rules governing [policies with racist outcomes like] use of force or prosecutorial discretion."[32] In fact, as Naomi Murakawa points out, focusing on accounts of individual bias risks embedding racism even further into our policies and practices by focusing on taking bias-prone decisions away from legal actors and cementing non-negotiable policies (e.g., mandatory minimums).[33,34] Thus, implicit bias

---

[26] Pew Research Center. (2016, June 27). On Views of Race and Inequality, Blacks and Whites Are Worlds Apart. *Pew Research Center's Social & Demographic Trends Project*. https://www.pewresearch.org/social-trends/2016/06/27/on-views-of-race-and-inequality-blacks-and-whites-are-worlds-apart/

[27] Rucker, J. M., & Richeson, J. A. (2021). Toward an understanding of structural racism: Implications for criminal justice. *Science*, *374*(6565), 286–290. https://doi.org/10.1126/science.abj7779

[28] Carter, E. R., Onyeador, I. N., & Lewis, N. A. Jr. (2020). Developing & delivering effective anti-bias training: Challenges & recommendations. *Behavioral Science & Policy*, *6*(1), 57–70. https://doi.org/10.1353/bsp.2020.0005

[29] Worden, R. E., McLean, S. J., Engel, R. S., Cochran, H., Corsaro, N., Reynolds, D., Najdowski, C. J., & Isaza, G. T. (2020). *The Impacts of Implicit Bias Awareness Training in the NYPD*. Center for Police Research and Policy, John F. Finn Institute for Public Safety, Inc. https://www1.nyc.gov/assets/nypd/downloads/pdf/analysis_and_planning/impacts-of-implicit-bias-awareness-training-in-%20the-nypd.pdf

[30] Dobbin, F., & Kalev, A. (2021, May 21). Why Diversity Training Does Not Work and Policies to Combat Bias in the Workplace More Effectively. *The Economist*. https://www.economist.com/by-invitation/2021/05/21/frank-dobbin-and-alexandra-kalev-explain-why-diversity-training-does-not-work

[31] Onyeador, I. N., Hudson, S. T. J., & Lewis, N. A. (2021). Moving Beyond Implicit Bias Training: Policy Insights for Increasing Organizational Diversity. *Policy Insights from the Behavioral and Brain Sciences*, *8*(1), 19–26. https://doi.org/10.1177/2372732220983840

[32] Muhammad, K. G. (2010). *The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America*.

[33] Murakawa, N. (2014). *The First Civil Right: How Liberals Built Prison America*.

[34] The Sentencing Project. (2018). *Report to the United Nations on Racial Disparities in*

science has the potential to reinforce the same structural and institutional manifestations of racism that they conceal.

If implicit bias science shifts our attention away from social policies and practices, who are we to blame for racism? Here, implicit bias science (misconstrued) offers us a convenient scapegoat: the brain. After all, implicit bias is *implicit* and therefore inaccessible, tempting us to believe that "*I'm* not responsible, *my brain* is!" (a distinction most brain scientists would be quick to dispel, but is nevertheless pervasive in society at large, as we will continue to see). Implicit bias science thus risks '*over-neuralizing'* racism, absolving us of (1) individual responsibility to intervene on our own biased behavior by deeming it inaccessible to us, and (2) collective responsibility to intervene on biased social conditions outside the scope of brain science. Over-neuralizing racism in this way leaves us disempowered in our own control over interpersonal *and* structural racism, and further might lull us into inaction.

To conclude, in an effort to address racism with the tools of brain science, psychology and neuroscience researchers developed the concept of implicit bias, which *by virtue of the methods used to define it* constrains racism to the scale of the individual brain. Political and social actors (e.g., police departments or discriminatory corporations) can then exploit this socially inattentive research to support an over-individualized account of racism as defined by implicit bias (e.g., 'bad apples')—one that neglects scientifically 'uncontrollable' aspects of racism and obscures system- and institution-level intervention on racial inequality.[35] Furthermore, implicit bias science risks over-neuralizing racism, offering a fallacious way to blame one's brain for racially discriminatory behaviors and alleviate responsibility for social action.

Brain scientists, including but not limited to implicit bias researchers, might have avoided these inferential leaps and inadvertent social consequences with the idea of 'embedded scholarship', an idea we will return to later, requiring researchers to take seriously how implicit bias research fits into a broader historical and sociological picture of systemic racism and how implicit bias research interfaces with public thought and policy.[35,36,37]

In the case of implicit bias, the translational scope of this research has been somewhat limited; however, in the case of addiction, the synergistic

---

the U.S. Criminal Justice System. https://www.sentencingproject.org/publications/un-report-on-racial-disparities/

[35] Rucker, J. M., & Richeson, J. A. (2021). Toward an understanding of structural racism: Implications for criminal justice. *Science*, *374*(6565), 286–290. https://doi.org/10.1126/science.abj7779

[36] Cogburn, C. D. (2019). Culture, Race, and Health: Implications for Racial Inequities and Population Health. *The Milbank Quarterly*, *97*(3), 736–761. https://doi.org/10.1111/1468-0009.12411

[37] Salter, P. S., Adams, G., & Perez, M. J. (2018). Racism in the Structure of Everyday Worlds: A Cultural-Psychological Perspective. *Current Directions in Psychological Science*, *27*(3), 150–155. https://doi.org/10.1177/0963721417724239

interplay of inattentive science with large-scale, discriminatory, and socially destructive policy initiatives becomes even clearer.

II. Addiction

In the past fifty years, neuroscience has become one of the preeminent ways of understanding drug addiction. The National Institute on Drug Abuse (NIDA) defines addiction as a "*brain disorder* instantiated in motivational and inhibitory systems, brought on by exposure to substances that pharmacologically impose lasting physiological changes on these systems" (emphasis added).[38] This 'brain disease' definition, however, risks over-neuralizing addiction and falsely centering brain science solutions as the optimal approach for understanding drug use and intervening on drug abuse.

To understand the motivation for the brain disease model of addiction, it is first necessary to describe its predecessor (and competitor): the 'moral model' of addiction.[39,40,41,42] In this view, supported by some psychologists and psychiatrists, addiction is simply a choice to use drugs to the point of severe psychological consequences. Proponents of the moral model argue that because no drug ensures addiction, those who become addicted must choose to do so. Thus, it casts drug addicts as morally deficient, for choosing to inflict the harms of addiction on themselves and those around them. The moral model *individualizes* addiction, already de-emphasizing its social determinants.

While the moral model of addiction raises noteworthy qualifiers about the addictive power of drugs in and of themselves (a point to revisit later), understanding the moral model is important to see the appeal of the brain disease model (a 'neuralized' alternative). By reimagining addiction as a disease of one's brain (in contrast to a moral defect), the brain disease model attempts to *humanize* addicted people by deeming them out of control. It shifts blame from the person to the drug, promoting social policy focused on eliminating or restricting drugs, as opposed to punishing addicted people.

However, the brain disease model falls prey to some of the same scientific inadequacies and social pitfalls described in the context of

---

[38] National Institute on Drug Abuse. (2018, July). *The Science of Drug Use and Addiction: The Basics*. National Institute on Drug Abuse. https://www.drugabuse.gov/publications/media-guide/science-drug-use-addiction-basics

[39] Heyman, G. M. (2009). *Addiction: A Disorder of Choice*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9xd9

[40] Pickard, H., Ahmed, S. H., & Foddy, B. (2015). Alternative Models of Addiction. *Frontiers in Psychiatry*, *6*, 20. https://doi.org/10.3389/fpsyt.2015.00020

[41] Satel, S., & Lilienfeld, S. (2014). Addiction and the Brain-Disease Fallacy. *Frontiers in Psychiatry*, *4*, 141. https://doi.org/10.3389/fpsyt.2013.00141

[42] Schaler, J. A. (2000). *Addiction is a choice*. Open Court.

racism. Scientifically, this model—largely championed by brain scientists—might overstate the neuroscientific dimensions of addiction. Opponents point out that no neurobiological marker has been identified to reliably differentiate an 'addicted brain' from a 'non-addicted brain' (as you might expect with other 'brain diseases,' like Alzheimer's).[43] Despite ostensible discoveries of 'brain damage' in addicted individuals, these researchers push back against the "disturbing tendency to interpret any brain *differences* as deficits representing substantial loss of brain function."[44] In fact, every aspect of ourselves—each desire, memory, or personality trait—are *in principle* observable on the brain, so the observability of brain differences between addicts and non-addicts does not warrant the ascription of a neurological disorder, especially absent clear evidence that these differences correlate to cognitive deficits.[43] Furthermore, in contrast to what the brain disease model might insinuate, the vast majority of drug users never become addicted (70-90% of those who use even the most stigmatized drugs, and higher for many other drugs), and most of those who do become addicted recover.[45,46,47]

In addition, the brain disease model fails to predict how addicted people behave. For much of the history of addiction science, researchers have suggested that drugs seize rational control from people and drive them to pursue their drug at any cost to themselves and people around them (e.g., 'compulsion theory' in James, 2007).[48] However, this irrationality does not hold up in animal or human multiple-choice studies, in which addicted subjects will repeatedly choose non-drug alternatives if given adequate alternatives and social conditions.[40,49,50] Environmental conditions, as opposed to neurological properties of the drug alone, exert an enormous influence on addiction-related behavior.

---

[43] Hart, C. L., Marvin, C. B., Silver, R., & Smith, E. E. (2012). Is Cognitive Functioning Impaired in Methamphetamine Users? A Critical Review. *Neuropsychopharmacology*, *37*(3), 586–608. https://doi.org/10.1038/npp.2011.276

[44] Hart, C. L. (2020). Exaggerating Harmful Drug Effects on the Brain Is Killing Black People. *Neuron*, *107*(2), 215–218. https://doi.org/10.1016/j.neuron.2020.06.019

[45]Hart, C. L. (2017). Viewing addiction as a brain disease promotes social injustice. *Nature Human Behaviour*, *1*(3), 1–1. https://doi.org/10.1038/s41562-017-0055

[46] Pickard, H. (2020). What We're Not Talking about When We Talk about Addiction. *Hastings Center Report*, *50*(4), 37–46. https://doi.org/10.1002/hast.1172

[47] Schlag, A. K. (2020). Percentages of problem drug use and their implications for policy making: A review of the literature. *Drug Science, Policy and Law*, *6*, 2050324520904540. https://doi.org/10.1177/2050324520904540

[48] James, W. (2007). *The Principles of Psychology*. Cosimo, Inc.

[49] Ahmed, S. H. (2010). Validation crisis in animal models of drug addiction: Beyond non-disordered drug use toward drug addiction. *Neuroscience and Biobehavioral Reviews*, *35*(2), 172–184. https://doi.org/10.1016/j.neubiorev.2010.04.005

[50] Venniro, M., Zhang, M., Caprioli, D., Hoots, J. K., Golden, S. A., Heins, C., Morales, M., Epstein, D. H., & Shaham, Y. (2018). Volitional social interaction prevents drug addiction in rat models. *Nature Neuroscience*, *21*(11), 1520–1529. https://doi.org/10.1038/s41593-018-0246-6

Finally, modern addiction neuroscience and the brain disease model have yielded remarkably little translational potential. As Pickard (2020) points out, most effective pharmacotherapies for opioid abuse (namely, methadone and buprenorphine treatment) were discovered as far back as the 1960s and 1970s, and the most effective treatments for cocaine abuse are based on behavioral principles that precede addiction neuroscience.[46] Taken together, these arguments do not reject any neurological aspect of addiction, but rather they suggest that it is not best understood (and treated) as a brain disease, per se.

Alternatively, these researchers propose a view of addiction as a highly heterogeneous, goal-directed pattern of drug use intricately related to social factors beyond the brain, and even beyond the individual. In the small percentage of drug users who become addicted—and to a greater extent, in the smaller percentage of addicted people who do not recover by their twenties or thirties—severe social adversity, co-occurring psychiatric disorders, and socioeconomic disempowerment account for a substantial amount of addiction[45,46] (however, the view of drug abuse as strictly a self-medication for other illnesses faces robust criticism).[51,52] Some of those who work directly with addicted people propose that there might be no universal characterization of addiction. Rather, addiction's heterogeneity (arising from unique life experiences, pressures, and triggers) should be reflected in addiction response.[46] As an (often overlooked) first step, "people need a 'a stake in conventional life': education, employment, housing, health, family, friends, community, belonging, respect, dignity, purpose, hope, self-worth, a sense of life's promise and possibility—the things that give life meaning and weigh heavily in the balance as a counter to the value of drugs."[46] The brain disease model obscures these non-neurological determinants of addiction, as well as non-neurological interventions that might follow.

As with implicit bias and racism, over-emphasizing neuroscientific (and therefore individual-level) accounts of addiction can be harmful, to both individuals and communities. For individuals, as previously noted, the brain disease model promotes unbacked ideas of irrationality and distorts public perception of the addictive power of drugs by focusing public attention solely on addicted users rather than (much more common) non-addicted ones. This distortion stigmatizes the possibility of healthy and productive drug use, subjects users to dangers of criminalization (e.g., police violence, incarceration), and stands in the way of honest drug education aimed at protecting and empowering users.[45]

For communities, as Carl Hart writes, "'neuro' remarks made about drugs with no foundation in evidence are pernicious: they help to shape an

---

[51] Lembke, A. (2012). Time to Abandon the Self-Medication Hypothesis in Patients with Psychiatric Disorders. *The American Journal of Drug and Alcohol Abuse*, *38*(6), 524–529. https://doi.org/10.3109/00952990.2012.694532

[52] Lembke, A. (2013). From self-medication to intoxication: Time for a paradigm shift. *Addiction*, *108*(4), 670–671. https://doi.org/10.1111/add.12028

environment in which there is an unwarranted and unrealistic goal of eliminating certain types of drug use at any cost to marginalized citizens."[45] In the United States, the emphasis of social policy on eliminating drugs (caused by the neuroscience-backed misconception that addictive power lies in the drug itself) has inflicted massive criminalization and harm—especially on low-income communities of color—through the 'War on Drugs.'[53] Further, criminalizing drugs has imposed barriers for addicted people to seek out medical treatment, mental health care, social support programs, and government-approved public information about the potential for harm from drug use.[53]

Moreover, the dissociation of one's brain and oneself ('my brain is responsible for me being addicted') again obscures productive interventions by offering the brain as a culprit for addiction. If we do not want to strictly blame the individual (as the moral model would have us do), shifting blame onto the brain saves us from accepting shared responsibility for meeting the social structural needs that seem to drive much addiction (e.g., socioeconomic disempowerment, racialized criminalization, inadequate mental health care). Here, we might draw upon addiction neuroscience to reject the moral model of addiction and explore the potential of neuropharmacological intervention, while contextualizing individual brains within broader social fabrics that the brain disease model is ill-equipped to account for.

Over-emphasizing neuroscience in research, policy, and media around drug addiction misconstrues addiction as a feature of one's brain, rather than a socially-rooted, injustice-driven phenomenon over which we all hold collective responsibility. The tools of brain science might discern neurological effects of drugs, but not phenomena like joblessness or inadequate mental health care. Thereby, addiction science—usually misguidedly, with disproportionate worry of drug abuse and neglect of safe drug use—risks distracting policymakers from structural components and absolving our responsibility to meet the social needs that fuel addiction, and for cultivating healthy drug use.

What is the way forward, in the case of addiction? Within science, I once more emphasize how embedded scholarship—earnestly engaging with social science and social policy surrounding socioeconomic drivers of addiction, racial biases in drug enforcement, and harms rendered by the criminalization of drugs—might cultivate an addiction neuroscience that is attentive and humble to the incomplete role of neuroscience in the addiction story, and responsive to the needs of both addicted and non-addicted users. This stance might point policymakers towards harm reduction approaches,[53,54] recovery-oriented care (rather than

[53] Earp, B. D., Lewis, J., & Hart, C. L. (2021). Racial Justice Requires Ending the War on Drugs. *The American Journal of Bioethics*, *21*(4), 4–19. https://doi.org/10.1080/15265161.2020.1861364

[54] Hoss, A. (2019). *Legalizing Harm Reduction* (SSRN Scholarly Paper ID 3439679).

criminalization and punishment),[55] honest drug education, and the regulated legalization of many drugs stigmatized by distorted neuroscientific narratives.[45]

III. Criminality

In the case of criminality, modern brain scientists continue a long (and sordid) legacy of misguidedly searching for the neural basis of crime, once again neglecting social complexities and scientific shortcomings with little regard for those enmeshed in the criminal justice system. As Michelle Alexander (2012) writes, "criminals are the one social group in America that nearly everyone—across political, racial, and class boundaries—feels free to hate." Furthermore, throughout the United States' history, the way crime is understood—especially in scientific and statistical contexts— has been racialized to target black and brown Americans [32,56,57] (see history of phrenology).[58] The concept of criminality, particularly in the United States, has long been immersed in fraught discourse about the overemphasis of retribution (as opposed to rehabilitation or restoration), the racialized nature of crime perception and response, and the dehumanization of those who commit crimes.

Brain scientists in the late twentieth century stepped into this discourse and began studying neuroscientific dimensions of criminal behavior with the intent of shifting focus from punishment to rehabilitation, and of obtaining objective, race-neutral biological markers to help do so. (Brain scientists had stepped into this discourse in the nineteenth century, and committed pseudoscientific and racist atrocities in the name of phrenology).[58] The tools of cognitive neuroscience, by enabling the explanatory connection between brain and behavior, opened the door for scientists to apply these tools to criminal behavior, with broad

Social Science Research Network. https://doi.org/10.2139/ssrn.3439679

[55] Humphreys, K., & Lembke, A. (2014). Recovery-oriented policy and care systems in the UK and USA. *Drug and Alcohol Review*, *33*(1), 13–18. https://doi.org/10.1111/dar.12092

[56] Alexander, M. (2012). *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. New Press.

[57] Roberts, D. (2018). The Ethics of Biosocial Science. *Tanner Lectures on Human Values*. https://scholarship.law.upenn.edu/faculty_scholarship/2505

[58] Thompson, C. E. (2021). *An Organ of Murder: Crime, Violence, and Phrenology in Nineteenth-Century America* (1st edition). Rutgers University Press.

implications for the criminal legal system.[59,60,61,62] These scientists envision neuroscience helping to replace punitive sentences with more consequentialist strategies of social management (e.g., intervening on mental illness, brain-based rehabilitative therapies). They note that in the worst case scenarios, where such rehabilitation is not feasible, neuroscience might improve our predictions of criminal dangerousness.[6] This neuro-legal theorizing, in turn, emerges from research on the '*criminal brain*'—a neuroscientific operationalization of what criminality and violent behavior looks like, in the brain.[63,64] Crucially, in this context, neuroscientific methods are characterized as an *objective* measure of crime because they are biological, neutral to race, class, gender, and other social dimensions.

In this section, I will discuss how the criminal brain in fact promotes the same racialized punitiveness it purports to avoid. Research on the criminal brain has operationalized criminal risk and behavior in an individual- and brain-focused manner that obscures situational and structural causes of crime, fails to yield the rehabilitative solutions it promises, and bolsters punitive measures of criminal dangerousness. Furthermore, by parading as an objective account of criminality, it has failed to engage with the socially-constructed nature of crime, and therefore risks reinscribing biologized and racialized understandings of crime. Lastly, we will see in the context of capital trials that it 'over-neuralizes' the concept of criminal culpability, reinforcing a fallacious dissociation between the brain and the self and absolving us of collective responsibility for the social conditions that produce violent behavior. Once more, the individual-focused perspectives of neuroscience falsely center the brain as the locus of understanding crime, distracting us from progress towards a more humane legal system.

To begin, the characterization of criminality as a brain trait faces the same scientific uncertainties and translational fruitlessness present in the

---

[59] Anderson, N. E., & Kiehl, K. A. (2020). Re-wiring Guilt: How Advancing Neuroscience Encourages Strategic Interventions Over Retributive Justice. *Frontiers in Psychology*, *11*, 390. https://doi.org/10.3389/fpsyg.2020.00390

[60] Eagleman, D. (2011, June 7). *The Brain on Trial*. The Atlantic. https://www.theatlantic.com/magazine/archive/2011/07/the-brain-on-trial/308520/

[61] Greene, E., & Cahill, B. S. (2012). Effects of Neuroimaging Evidence on Mock Juror Decision Making. *Behavioral Sciences & the Law*, *30*(3), 280–296. https://doi.org/10.1002/bsl.1993

[62] Zeki, S., Goodenough, O. R., & Sapolsky, R. M. (2004). The frontal cortex and the criminal justice system. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1451), 1787–1796. https://doi.org/10.1098/rstb.2004.1547

[63] Raine, A. (2013). *The Anatomy of Violence: The Biological Roots of Crime* (Illustrated edition). Vintage.

[64] Rollins, O. (2021a). *Conviction: The Making and Unmaking of the Violent Brain*. Stanford University Press.

cases of implicit bias and addiction. Neuroscientific findings about violence as a proxy for crime, tend to resemble differences in brain structure and function between participants deemed 'criminal' (often by certain psychiatric diagnoses, or just violent criminal records) and participants deemed 'not-criminal.' For example, Adrian Raine and colleagues (2000) compared men diagnosed with Antisocial Personality Disorder (ASPD) to non-ASPD controls and found in the former group volumetric deficits in the prefrontal cortex (a brain region commonly associated with decision-making and executive control over one's behavior).[65]

Such studies, however, fall well short of their stated goals. First, findings like the one above often draw upon sample sizes and statistical practices now thought to be egregious, with unclear replicability.[66] Second, as we saw in the case of addiction, these claims specify brain *differences* in certain criminalized groups (here, people diagnosed with ASPD); the normative leap from 'different' to 'damaged' is arbitrarily ascribed by the researcher.[66] Third, such neuroimaging research cannot tell us the cause of such brain differences, nor if they play any causal role in violent behavior. Volumetric brain differences could, for instance, be caused by the conditions of prison confinement themselves (or any number of complex social inputs unmeasured in these studies), rather than being directly related to violent behavior. Fourth, much of the weight of these claims rests upon operationalizations of 'criminal' that are potentially subjective, socially complex, systematically biased, or even self-fulfilling, meriting strong caution interpreting these findings as objective.[67,68,34] Fifth, it is worth considering the interventional utility of such studies. Besides a few spare speculative ideas (e.g., neurofeedback therapy),[60] neurally localizing violence within the brain has not yielded the promised rehabilitative interventions of the neuroscientific justice endeavor (and in fact, solutions truly enabled by this research might resemble advanced forms of lobotomy or other phrenology-era solutions). Finally, as we have seen before, searching for the root of violence in the brain shifts our gaze away from social conditions of marginalization and disempowerment that might be equally or more causative of violence than

---

[65] Raine, A., Lencz, T., Bihrle, S., LaCasse, L., & Colletti, P. (2000). Reduced prefrontal gray matter volume and reduced autonomic activity in antisocial personality disorder. *Archives of General Psychiatry*, *57*(2), 119–127; discussion 128-129. https://doi.org/10.1001/archpsyc.57.2.119

[66] Rollins, O. (2018). Risky Bodies: Race and the Science of Crime. In T. Rajack-Tally & D. Brooms (Eds.), *Living Racism: Through the Barrel of thee Book* (pp. 91–119). Lexington Press.

[67] Cunningham, M. D., & Reidy, T. J. (1998). Antisocial personality disorder and psychopathy: Diagnostic dilemmas in classifying patterns of antisocial behavior in sentencing evaluations. *Behavioral Sciences & the Law*, *16*(3), 333–351. https://doi.org/10.1002/(SICI)1099-0798(199822)16:3<333::AID-BSL314>3.0.CO;2-N

[68] Szasz, T. S. (1960). The myth of mental illness. *American Psychologist*, *15*(2), 113–118. https://doi.org/10.1037/h0046535

personal characteristics.[32,64] While violent behaviors—like all behaviors—are visible in the brain, this does not mean the root causes of violence lie within (or are best understood in the context of) the brain.[66]

In the absence of scientific clarity and rehabilitative utility, the idea of the criminal brain has been exploited for what was mentioned as 'the worst case' scenario above: calculating criminal risk. Already, neurobiological correlates of violent behaviors are being used in forensic assessments of criminal risk (for review),[69] an aspect of the legal system that often perpetuates more punitive and racially disparate responses.[70] In the closely related field of neurogenetics, one defendant (Jeffrey Landrigan) who offered evidence of genetic predispositions towards violence to hopefully alleviate his criminal responsibility was sentenced to death precisely on the intuition that his neurogenetic abnormalities signaled the pervasiveness of criminality in his psyche and the impossibility of rehabilitation.[71] Neuroscience research on the criminal brain has failed to guide us towards rehabilitative interventions, and in their absence, it has been drawn upon to sharpen punitive aspects of the criminal legal system.

Furthermore, neuroscientific operationalization of criminality has taken for granted the normative structure of the criminal legal system, and has perpetuated misunderstandings of crime in several key ways. As noted above, it often uses violence and aggression as a proxy for crime, heedlessly accepting the equivalence of the two without acknowledging that only certain forms of violence are criminalized by the legal system and many criminal acts are not violent at all.[64] As a particularly salient example, acquitted perpetrators of unwarranted police violence would not be studied in these experiments. Especially with the often race- and class-biased definitions and enforcement of crime, failing to engage with this normative assumption of the legal system risks perpetuating unequal or unjust enforcements of crime.[56]

In addition, in similar fashion to neuroscience of implicit bias, neuroscience of criminality underestimates the pervasiveness of racialization in our environments, which, combined with its claim of objectivity, risks reinscribing racialized notions of criminality onto the brain. When studying neural aspects of crime, researchers must choose which brains are 'criminal,' and therefore the object of study. Using race-biased indices (e.g., who tends to be arrested by the police and charged with violent crimes, or even who tends to commit violent crime) thus embeds racialization into the 'objective' measurement of criminality that

---

[69] Gronde, T. van der, Kempes, M., El, C. van, Rinne, T., & Pieters, T. (2014). Neurobiological Correlates in Forensic Assessment: A Systematic Review. *PLOS ONE*, *9*(10), e110672. https://doi.org/10.1371/journal.pone.0110672

[70] Green, B. (2020). The false promise of risk assessments: Epistemic reform and the limits of fairness. *FAT**. https://doi.org/10.1145/3351095.3372869

[71] Jones, O. D. (2006). Behavioral Genetics and Crime, in Context. *Law and Contemporary Problems*, *69*(1/2), 81–100.

results from such research. Once again, the individual-level focus of neuroscience sidesteps the tenacity and complicated functioning of racialization in society and takes as *race-neutral* the highly *racialized* environments in which the brains and individuals of study are situated. Doing so risks perpetuating a long history of using objective-seeming biological science to inscribe certain racialized groups (predominantly, black Americans) as 'criminal.'[57,66] In other words, if we treat crime as an objective, biological trait, then we biologize the race-based roots of how crime is caused, perceived, and defined.

Thus, the neuroscientific operationalization of criminality (the criminal brain), while professing rehabilitative goals and race-neutral measures, has sustained both the punitive nature of the criminal legal system and a (normatively, scientifically, and socially) incomplete picture of what criminal risk is. The ramifications of these shortcomings go beyond the abstract theorizing here; courtroom decisions of criminal *culpability* draw upon the idea of the criminal brain, rendering very real consequences to trial decisions.

In the early 2010s, neuroscientific evidence was introduced in 25% of capital trials in the United States.[72] Such evidence was usually proffered during the trial's sentencing phase, in which a defendant has already been found guilty and the judge now decides their sentence (frequently, in such cases, between life imprisonment and the death penalty). Specifically, neuroscientific evidence was usually *mitigating evidence*, intended to diminish a defendant's criminal culpability with proof of brain abnormalities or neurological damage that might have played a causal role in one's criminal behavior (i.e., shifts criminal culpability from *oneself* to *one's brain*, similarly to how the 'rotten social background defense' shifts culpability onto one's environments). Importantly, such neuroscientific evidence seems to prove convincing of less punitive sentences; participants in mock jury studies tend to at least forego the death penalty when documentation of 'neuroscientific deficits' is forthcoming. [61,73,74]

The influence of brain science on capital punishment decisions also extends into Supreme Court jurisdiction. For example, in *Roper v. Simmons*, neuroscientific research cited in amicus briefs from the American Psychological Association and American Medical Association were especially formative in abolishing the death penalty for minors.

---

[72] Farahany, N. A. (2015). Neuroscience and behavioral genetics in US criminal law: An empirical analysis. *Journal of Law and the Biosciences*, *2*(3), 485–509. https://doi.org/10.1093/jlb/lsv059

[73]Appelbaum, P. S., Scurich, N., & Raad, R. (2015). Effects of Behavioral Genetic Evidence on Perceptions of Criminal Responsibility and Appropriate Punishment. *Psychology, Public Policy, and Law: An Official Law Review of the University of Arizona College of Law and the University of Miami School of Law*, *21*(2), 134–144. https://doi.org/10.1037/law0000039

[74] Saks, M. J., Schweitzer, N. J., Aharoni, E., & Kiehl, K. A. (2014). The Impact of Neuroimages in the Sentencing Phase of Capital Trials. *Journal of Empirical Legal Studies*, *11*(1), 105–131. https://doi.org/10.1111/jels.12036

These briefs cited studies documenting impaired executive functioning and still-developing prefrontal cortices in adolescents, suggesting the death penalty was 'cruel and unusual' given that impulsive adolescent behavior could be traced back to neuropsychological differences between adolescents and adults (for other relevant examples, see *Ford v. Wainwright* and *Atkins v. Virginia*).[75,76] These cases, taken together, codify the same intuition as seen above—that tracing criminal behavior back to observable brain differences diminishes one's criminal culpability, shifting blame from oneself to one's brain.

The use of neuroscientific evidence here over-neuralizes a defendant's criminal culpability; however, because of the quite literal life-or-death stakes, its value should not be denied. Sometimes, in the example of an easily excisable brain tumor driving aggressive behavior, this intuition is helpful, and suggestive of remedy. In other cases, a neuroscientific finding might bolster claims about *behaviors* or *environmental factors* (that is, non-neural factors)—for example, the malleability of youthful recklessness, or the long-lasting effects of a previous trauma that might inspire a more merciful sentence.

However, deeming someone more or less culpable because of the presence or absence of a neuroscientific finding itself—independent of suggested remedy or as evidence of another mitigating factor—can belie a fundamental notion of cognitive neuroscience. In principle, *all* behavior can be traced to some pattern of connection or activity in the brain.[1] Our current neuroscientific tools are only powerful (spatially and temporally specific) enough to document certain behaviors, thoughts, and cognitive events, but in theory, we will one day be able to view the neural signature of any behavior on the brain.[60] The intuition that neuroscientific measurement, per se, alleviates criminal culpability, then, rests one's moral responsibility (and consequent culpability) on the state of neuroscientific tools (which will continue to improve, and thus shrink the scope of culpable behaviors). The true power of such neuroscientific findings is in illuminating mental, social, and environmental conditions that drove violent behavior. Absent such considerations, shifting blame from the individual (the criminal legal status quo) to the brain (the neuro-legal alternative) again obscures structural and social drivers of violence—and supports a court's neglect of those types of evidence. Blaming the brain allows us to show a level of mercy towards *certain* defendants who meet a misleading standard of neural culpability without accepting collective responsibility for these social conditions that spawn crime.

---

[75] Cauffman, E., & Steinberg, L. (2000). (Im)maturity of judgment in adolescence: Why adolescents may be less culpable than adults*. *Behavioral Sciences & the Law*, *18*(6), 741–760. https://doi.org/10.1002/bsl.416

[76] Smith, D. G., Xiao, L., & Bechara, A. (2012). Decision making in children and adolescents: Impaired Iowa Gambling Task performance in early adolescence. *Developmental Psychology*, *48*(4), 1180–1187. https://doi.org/10.1037/a0026342

Taken together, these two instances of the criminal brain at work—to 'neuralize' criminal behavior and risk, and to 'neuralize' criminal culpability—fall into the same pitfalls that we have encountered throughout this paper. By oversimplifying violence and criminality to the level of the brain (the level on which neuroscientific tools become useful), neuroscience crowds out structural and environmental explanations of crime, and encourages us to understand violence in a vacuum—detached from the social conditions that shape our brains in the first place, and ultimately only useful for refining instruments of punitiveness. Further, the idea of the criminal brain reinscribes racialized and hierarchical inequalities in our society into biologized brain traits, perpetuating a sordid history of seeking an 'objective' biological rationale for racializing what a 'criminal' is. Lastly, in practice, it again promotes the idea that one's brain is dissociable from oneself, allowing us to shift blame for crime from a person to a brain, all the while neglecting social conditions that breed violence—the very same conditions that confound the tools of neuroscience and are therefore excluded from neuroscientific study of criminality. As we saw first with implicit bias, and then with addiction, the excitement of neuroscientists and legal actors alike over the potential of neuroscience to revolutionize criminal justice over-emphasized neural explanations for crime, often at a high cost to those involved in the criminal legal system.

Conclusion

In the cases of racism, addiction, and criminality, brain scientists have sought to apply neuroscientific insights for social progress and justice. Doing so has produced the concepts of the 'biased brain,' the 'addicted brain,' and the 'criminal brain,'—over-individualized and over-neuralized views of these complex social phenomena that distract from structural barriers and absolve collective responsibility for addressing those barriers.

In each of these cases, the consequences of socially incomplete research and misguided scientific translation have spanned from individuals—racially biased interactions, stigmatized drug users, and victimized defendants—all the way to entire institutions—police departments, drug wars, and courtrooms. Brain scientists have often put racism, addiction, and crime in neural vacuums, neglecting historical and material complexities of our society that contribute to those issues, but fall outside the purview of brain science. Moreover, the emergence of the 'social brain' offered a new culprit for these injustices, allowing us to keep our focus on the brains of disadvantaged individuals, and even show some mercy towards those individuals, without accepting collective responsibility for underlying social conditions of disadvantage. These dangers, it should be noted, extend beyond the three themes discussed here—for example, to certain aspects of education, mental health, and economic decision making, to name a few.

This is not to condemn the application of brain science to social justice, or to dispute the revelatory potential of doing so. In fact, we should not ignore the individual or the brain in favor of only the social conditions, either. Accepting collective responsibility for social conditions of injustice has been my focus here, to illuminate powerful perspectives and interventional strategies that centering brain science has often concealed. However, my aim is not to shift all responsibility for racism, addiction, and crime from the individual (and brain) onto the social environment; rather, it is to show the shortcomings in our rhetorical division between the two. Cognitive neuroscience and psychology have shown us that the brain—and therefore, the self— are intricately entangled with our social environments. Our memory systems encode systematically stereotyped information from the media and recapitulate biases in our behavior. Our most fundamental and often insurmountable cravings are shaped by our economic, political, and social conditions. The violence that we commit to each other as individuals is inextricable from the violence that we commit on communities, on subordinated groups, and on entire regions of the world. Individual and collective responsibility are continuous, and my aim here is not to alleviate all individual responsibility for racism, addiction, and crime, but rather to open us to the possibility of a different kind of individual responsibility: one that illuminates us to the shortcomings of our social structures and environments and primes us to take seriously individual- and brain- level harms and needs.

To this end, brain science might certainly prove useful, and even transformative. However, to apply these insights to social justice, we have seen the array of scientific and social pitfalls that stand in the way of socially valuable brain science. To cultivate such research, I offer three recommendations:

1. I again underscore the idea of 'embedded' and humble neuroscientific scholarship—a manner of research that requires engagement with what social science and critical theory reveal about the conditions in which our brains are situated, to highlight where claims about the individual brain will not tell the whole story, and to avoid 'controlling away' (and therefore, as we have seen, reinscribing onto the brain as biologized traits) the complexities of racialization and social injustice that characterize our society.[35,36,37]
2. Beyond simply knowing when individual claims fall short, the case of implicit bias research has shown us that brain science might indeed reveal traits of environments and social structures, or at least cognitive expressions of them. Findings like these need not direct us towards individual-level interventions like anti-bias coping strategies; rather, they can direct us towards eliminating structural causes and drawing hope from neural plasticity that removing structural barriers may ameliorate a harmful effect.[57]

Being open to making claims beyond the individual requires dutiful attention to inter- and intra-subject variability, to clarify the explanatory level of neuroscientific findings.

3. Lastly, brain scientists must be resolute in having a stake in the communication and translation of science beyond the laboratory.[77] In the case of addiction, many of the misconceptions at the heart of the War on Drugs are enabled by poor communication of what the 'brain disease' model truly entails, with disastrous and often fatal consequences for many victims of drug criminalization. This endeavor requires attention to the aforementioned social causes of a particular biological phenomenon, and commitment to solving these issues at their root rather than focusing on intervening on those suffering from their effects.[57] This endeavor also requires being clear and honest about the translational potential of research—if localizing violence in the brain will not yield restorative changes to criminal justice, clarity of that fact might avoid the brazen recruitment of neuroscience for punitive ends.

Taken together, these recommendations encourage us to embrace a complex and often uncontrollable picture of the ways that social injustice shapes our brains and inserts itself into neuroscientific inquiry, such that we might produce research truly responsive to injustice, informed by the complexity of human experience and aimed towards clear and conscientious social change.

[77] IJzerman, H., Lewis, N. A., Przybylski, A. K., Weinstein, N., DeBruine, L., Ritchie, S. J., Vazire, S., Forscher, P. S., Morey, R. D., Ivory, J. D., & Anvari, F. (2020). Use caution when applying behavioural science to policy. *Nature Human Behaviour*, *4*(11), 1092–1094. https://doi.org/10.1038/s41562-020-00990-w