NEUROLAW: THE CALL FOR ADJUSTING THEORY BASED ON SCIENTIFIC RESULTS

EDITED BY: José M. Muñoz, Eric García-López and Elena Rusconi

PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-208-1
DOI 10.3389/978-2-88966-208-1

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

NEUROLAW: THE CALL FOR ADJUSTING THEORY BASED ON SCIENTIFIC RESULTS

Topic Editors:
José M. Muñoz, Universidad Europea de Valencia, Spain
Eric García-López, Instituto Nacional de Ciencias Penales, Mexico
Elena Rusconi, University of Trento, Italy

Image: iStock.com/erhui1979

Acknowledgments
N.B. This Research Topic idea was co-developed with Dr. Nicolás Ezequiel Llamas, Dr. José Ángel Marinaro and Dr. Ezequiel Norberto Mercurio.

Citation: Muñoz, J. M., García-López, E., Rusconi, E., eds. (2020). Neurolaw: The Call for Adjusting Theory Based on Scientific Results. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-208-1

Table of Contents

frontiers in Psychology

EDITORIAL
published: 09 October 2020
doi: 10.3389/fpsyg.2020.582302

Editorial: Neurolaw: The Call for Adjusting Theory Based on Scientific Results

José M. Muñoz[1], *Eric García-López*[2*] *and Elena Rusconi*[3]

[1] Department of Psychology, Universidad Europea de Valencia, Valencia, Spain, [2] Instituto Nacional de Ciencias Penales, Mexico City, Mexico, [3] Department of Psychology and Cognitive Science, University of Trento, Trento, Italy

Keywords: neurolaw, free will, criminal responsibility, psychopathy, neuroprediction, adolescent brain, addiction, decision neuroscience

Editorial on the Research Topic

Neurolaw: The Call for Adjusting Theory Based on Scientific Results

The Research Topic (RT) presented here is about the complex relationship between Law and Neuroscience. We hope that it strengthens the dialogue between both disciplines across the globe, not only for a better understanding of human behavior in the legal and forensic context, but also for a better comprehension about the meaning of Justice with a view from Neuroscience. From this collection, very different positions emerge on the possible use of neuroscientific evidence to inform the law and criminal justice interventions (e.g., to safeguard personal freedom and dignity vs. to exert social control). This RT has been written by researchers from leading universities around the world and all of the papers included are based on scientific results and the most relevant and up-to-date information in each topic.

In Part I of this RT (*Neuroscience and the law: Can we fit them together?*), Pernu and Elzein advocate separating different neural-based perspectives according to their degree of viability to inform moral and legal debates on human decisions and actions, and argue that a view of neurolaw based on mixing the various perspectives is a reason for neuroscience not having strongly permeated the law as yet. Bigenwald and Chambon warn us about the possibility of a revolution in Criminal Responsibility. "Not yet" they say, by explaining what is called "the limits of Neuroscience" in their article. These limits are not only technical but legal: "[....] neuroscience can only impact legal excuses and not legal justifications." They conclude that: "While neurolaw often evokes the neuroscientification of law, it could more properly refer to the juridification of neuroscience, i.e., legal thinking that would integrate and apply scientific discoveries to criminal justice." Anderson and Kiehl, for their part, hold that neuroscience favors a change in normative attitudes—moving away from retributivist approaches—that, far from radically modifying the process of the legal assignment of guilt, allows it to be improved by adopting "more pragmatic strategies for combating the most conspicuous patterns promoting mass incarceration and recidivism."

In Part II (*Neurolaw and psychopathy*), van Dongen writes about a "crucial human ability" (empathy) and focuses on the "social brain of psychopaths." She argues that we must work on the "elucidation of the neural underpinnings of empathy" and then that we should think about "neurophysiological informed personalized treatment interventions that ultimately reduce violent transgressions in individuals with psychopathic traits." Also, she brings an overview about psychopathy and a bio-cognitive perspective for such disorders. This second part continues with a commentary on an article by Baccarini and Malatesti (2017) in which they advocate a non-consensual application of moral bioenhancement—gene editing, neurosurgery, psychotropic treatment, etc.—to psychopaths. Their position is based on maintaining that a psychopath would

OPEN ACCESS

Edited by:
Chiara Fini,
Sapienza University of Rome, Italy

Reviewed by:
Andrea Lavazza,
Centro Universitario Internazionale, Italy
Toma Strle,
University of Ljubljana, Slovenia

***Correspondence:**
Eric García-López
garcialopez@gmx.com

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *11 July 2020*
Accepted: *28 August 2020*
Published: *09 October 2020*

Citation:
Muñoz JM, García-López E and Rusconi E (2020) Editorial: Neurolaw: The Call for Adjusting Theory Based on Scientific Results. Front. Psychol. 11:582302. doi: 10.3389/fpsyg.2020.582302

allow that moral bioenhancement be applied to other psychopaths and therefore she must be treated the same way. In the commentary, Sirgiovanni and Garasic give reasons against this argument, believing that non-consensual treatment is unjustified in this case, and ultimately holding that such an invasive treatment as moral bioenhancement must be consented by the psychopath. Guillen Gonzalez et al. address the impact of biological evidence on sentencing decisions for psychopathic offenders. In a sample of German law students, evidence of brain injury but not of genetic predisposition lowered legal responsibility judgments compared to when no biological evidence was provided by the defense. No effects were found on the length of sentencing, similar to a previous study on German judges (Fuss et al., 2015) and unlike a previous study on U.S. judges (Aspinwall et al., 2012) where genetic predisposition caused the assigned prison sentence to be lowered. The authors argue that differences in criminal justice systems may explain the differential effects of biological evidence.

Tortora et al. begin Part III (*Recent advances in risk assessment*) by analyzing current evidence about how brain-reading technology—a product of the convergence between neuroimaging and AI—could be applied to forensic psychiatry and criminal justice as a tool for risk assessment and neuroprediction of violence and future recidivism. They conclude that further research must be done in this regard, and also that we would do well to anticipate debates about benefits and damages of these eventual applications of brain-reading. Haarsma et al. bring us the results of testing probationers in Houston, TX from 2017 to 2019 with a mobile neurocognitive software to predict reoffense. This NeuroCognitive Risk Assessment (NCRA) "opens the possibility of identifying different levels of recidivism risk, by crime type, for any age, or gender, and seeks to steer individuals appropriately toward rehabilitative programs."

In Part IV (*Neuroscience and adolescent legal responsibility: The Latin American case*), Mercurio et al. address the importance of Neurobiology for the age of criminal responsibility. They argue that there is no scientific evidence to reduce the age of criminal punishment and they are "disposed not to recommend lowering the age of criminal responsibility, but rather increasing it." This article reminds us that "Latin America does not benefit enough from the advances of the neuroscience in its application to legal issues" (García-López et al., 2019, p. 14), and also that we need that all these countries become part of this new perspective. Llamas and Marinaro draw attention to the existence of a wide range of legislative methods across Latin America concerning juvenile justice. They highlight how some of those methods may be at odds with international law and do not take into account a growing body of neuroscientific evidence showing important differences between adolescent and adult brain functioning. They advocate a revision of penological justifications in those judicial systems that still allow the application of similar punishments to juvenile offenders and adult offenders for the same crime.

Part V (*Special topics*) includes contributions on two issues that are not regularly addressed in neurolaw mainstream debates but which are relevant to make this discipline more comprehensive: moral responsibility in cases of addiction, and decision neuroscience in relation to customers' choices. Rise and Halkjelsvik present a triple study on how the different ways in which people conceive addiction cause different moral judgments when it comes to attributing responsibility to an agent. Their study showed that this attribution was "lower when addiction was connected to diseases and disorders, such as dysfunctional processes in the brain, and greater when addiction was associated with agency and addictive behaviors." Bault and Rusconi draw attention to neuroscientifically-informed techniques that are currently used in marketing to manipulate consumer behavior and how certain groups may be particularly vulnerable to such manipulations. The growing efficacy of these techniques calls for regulatory interventions that are not limited to specific products (e.g., sweets and cigarettes) and age groups (e.g., children) but take into account our understanding of the brain circuitry for decision and of vulnerabilities to external influences, in order to preserve freedom of choice.

This RT is composed by an up-to-date collection that is both comprehensive in terms of the topics covered and balanced with regard to the inclusion of empirical, analytical, and review studies. We believe these characteristics make it a valuable tool equally useful to those interested in an introduction on neurolaw and to those looking to keep up to date with the latest and most innovative research in this fascinating and emerging discipline.

AUTHOR CONTRIBUTIONS

All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank our Argentine colleagues Dr. Nicolás Ezequiel Llamas, Dr. José Ángel Marinaro, and Dr. Ezequiel Norberto Mercurio for their valuable help in co-developing this RT idea. We also thank our Italian colleagues Dr. Marco Tullio Liuzza and Dr. Daniela Smirni for their work as handling editors for three manuscripts in this RT. Last, but not least, we thank all the reviewers for their valuable work on all of the papers included in this RT. This work is essential to make science progress. JM and EG-L state that their contribution is framed within the following research projects.

- *Espacios de progresión de las neurociencias en el derecho: aplicación en el campo de los derechos humanos, derecho penal, ejecución de la pena, neurociencia forense y neurotecnologías*, Universidad Nacional de La Matanza (PROINCE Program, Argentine Ministry of Education). Principal Investigator: Prof. José Ángel Marinaro.
- *Neuroderecho y psicopatología forense*, Instituto Nacional de Ciencias Penales. Principal Investigator: Prof. Eric García-López.
- *Derecho penal y comportamiento humano* (MICINN-RTI2018-097838-B-100), granted by the Spanish Ministry of Science, Innovation, and Universities (MCIU/AEI/FEDER, UE), program: "I+D+i orientada a los retos de la sociedad." Principal Investigator: Prof. Eduardo Demetrio Crespo. Website: *https://blog.uclm.es/proyectodpch/*

REFERENCES

Aspinwall, L. G., Brown, T. R., and Tabery, J. (2012). The double-edged sword: does biomechanism increase or decrease judges' sentencing of psychopaths? *Science* 337, 846–849. doi: 10.1126/science.1219569

Baccarini, E., and Malatesti, L. (2017). The moral bioenhancement of psychopaths. *J. Med. Ethics* 43, 697–701. doi: 10.1136/medethics-2016-103537

Fuss, J., Dressing, H., and Briken, P. (2015). Neurogenetic evidence in the courtroom: a randomized control trial with German judges. *J. Med. Genet.* 52, 730–737. doi: 10.1136/jmedgenet-2015-103284

García-López, E., Mercurio, E., Nijdam-Jones, A., Morales, L. A., and Rosenfeld, B. (2019). Neurolaw in Latin America: current status and challenges. *Int. J. Forensic Ment. Health* 18, 260–280. doi: 10.1080/14999013.2018.1552634

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 22 October 2020
doi: 10.3389/fpsyg.2020.01862

From Neuroscience to Law: Bridging the Gap

Tuomas K. Pernu[1,2]* *and Nadine Elzein*[3]

[1] Helsinki Collegium for Advanced Studies, University of Helsinki, Helsinki, Finland, [2] Department of Philosophy, King's College London, London, United Kingdom, [3] University of Oxford, Lady Margaret Hall, Oxford, United Kingdom

Since our moral and legal judgments are focused on our decisions and actions, one would expect information about the neural underpinnings of human decision-making and action-production to have a significant bearing on those judgments. However, despite the wealth of empirical data, and the public attention it has attracted in the past few decades, the results of neuroscientific research have had relatively little influence on legal practice. It is here argued that this is due, at least partly, to the discussion on the relationship of the neurosciences and law mixing up a number of separate issues that have different relevance on our moral and legal judgments. The approach here is hierarchical; more and less feasible ways in which neuroscientific data could inform such judgments are separated from each other. The neurosciences and other physical views on human behavior and decision-making do have the potential to have an impact on our legal reasoning. However, this happens in various different ways, and too often appeal to any neural data is assumed to be automatically relevant to shaping our moral and legal judgments. Our physicalist intuitions easily favor neural-level explanations to mental-level ones. But even if you were to subscribe to some reductionist variant of physicalism, it would not follow that all neural data should be automatically relevant to our moral and legal reasoning. However, the neurosciences can give us indirect evidence for reductive physicalism, which can then lead us to challenge the very idea of free will. Such a development can, ultimately, also have repercussions on law and legal practice.

Keywords: agency, causation, culpability, free will, liability, methodological dualism, neurolaw, prefrontal cortex

OPEN ACCESS

Edited by:
Marco Tullio Liuzza, Magna Graecia University of Catanzaro, Italy

Reviewed by:
Federica Nanci, Magna Graecia University of Catanzaro, Italy
Cristina Sanchez-Castañeda, University of Barcelona, Spain
Federico Gustavo Pizzetti, University of Milan, Italy

***Correspondence:**
Tuomas K. Pernu
tuomas.pernu@helsinki.fi

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *10 January 2020*
Accepted: *07 July 2020*
Published: *22 October 2020*

Citation:
Pernu TK and Elzein N (2020) From Neuroscience to Law: Bridging the Gap. Front. Psychol. 11:1862. doi: 10.3389/fpsyg.2020.01862

INTRODUCTION

According to a naturalistic, scientific, world view, reality is ultimately physical. Therefore, the human mind – our decision-making and behavior – must also be fundamentally physical. It would seem to follow, then, that the neurosciences, that study the physical basis of our minds, should be directly useful in understanding human decision-making and behavior, and should therefore also inform our moral and legal judgments.

Although this line of thinking is basically correct, it is all but clear how, exactly, neuroscientific evidence should bear on our moral and legal judgments. Here we outline a way of getting clearer on this by putting the question on the relevance of neuroscientific evidence to moral and legal reasoning in the more general context of metaphysics and the philosophy of science. Efforts to incorporate neuroscientific data into legal proceedings have had, at best, a mixed reception. We argue that much of the difficulty associated with the efforts to incorporate neuroscientific evidence in legal practice comes from a deeper problem of reconciling two radically different perspectives:

ontological monism that pervades our scientifically based thinking about the metaphysics of mind, and methodological dualism that governs our folk psychological reasoning, and which cannot easily be eliminated within the practical constraints of legal contexts. While it is a mistake to suppose that neuroscientific data is wholly irrelevant to jurisprudence, or that it cannot in some cases help to determine legal responsibility, we need to exercise caution in attributing responsibility on the basis of such data. At worst, those drawing on such evidence in order to undermine claims of moral and legal responsibility might be accused of trading on unwarranted interactionist assumptions, where these involve a conflation of neural realizers of mental states with external causes of them. However, we argue that such cases of bad neuroscientific reasoning should not obscure the value of neuroscientific evidence in other cases. In particular, we need to make a distinction between changes in neural features that might plausibly be described as involving natural rewiring in the brain, and changes that we have adequate and independent grounds for classifying as involving external interferences to ordinary brain function. Here, we survey the way in which neuroscientific evidence has come to be increasingly utilized in legal contexts, evaluating the different ways in which such evidence is presented with the above distinction in mind. We highlight three different ways in which the neurosciences can, or cannot, be used to inform our moral and legal judgments. We think that the discussion on neuroscience and law has been conflating these issues, which explains why neuroscientific evidence has received a varied response in legal practice.

First, it seems that there is some quite obviously bad reasoning often done on the basis of neuroscientific evidence (see section on "Lessons From Physicalistic Monism and Methodological Dualism" below). It should be clear that just pointing to *some* neuroscientific data is not evidence of these neural correlates being the source of, or even relevant to, a given mental or behavioral phenomenon: we already know that brain-functioning is necessary for all mental and behavioral phenomena, and to assume otherwise would amount to committing a dualistic fallacy – the fallacy, in this case, of inferring the irrelevance of psychological notions on the sole basis of pointing to their neural correlates (cf. Pernu, 2011; Elzein, 2019). So, simply noting that there are some (homogeneous) neural correlates of the ways of behaving we deem immoral or illegal should not make one think that those correlates are causing that sort of behavior [cf. e.g., Glannon (2011) and Morse (2011a, 2015) in relation to discussion in section on "Lessons From Physicalistic Monism and Methodological Dualism" below].

Second, there are also better, and to at least some extent valid, ways of taking the neuroscientific evidence into account in our moral and legal reasoning (as discussed in section on "Basing Lack of Agential Control on Neuroscientific Data" below). This can, in principle at least, be done by first separating different mental faculties' bearing on agential control from each other, and then showing that the functioning of some of the components essential for exercising those faculties has become dysfunctional for biological reasons. More precisely, in some cases we may be able to construct convincing evidence that there was some threat to agential control present due to neural factors on the basis that we have some independent evidence for a lack of control, and we can then point to a neural correlate for such a lack of control [e.g., Burns and Swerdlow (2003), **Box 7** below]. Establishing such connections is practically very difficult, and we still have a lot to learn about the psychology and neuroscience of agential control, but there are no principled reasons why such connections could not be established.

Third, and contrary to some intuitions stemming from physicalist metaphysics, neuroscience cannot, by itself, disprove the ideas of agency and free will (as discussed in section on "Physicalism, Free Will, and Moral Responsibility" below). In cases where moral or legal judgments are based on neural evidence the conclusions follow precisely because we can compare cases of lack of control to normal control cases, and point to their neural differences (and maybe abnormalities). No such contrast can be made in more global worries concerning agency and free will, for we are not able to compare cases where free will is exercised to cases where it is not. There is, in other words, an often-neglected difference between establishing exculpating factors in a particular legal case, and appealing to neuroscientific data that would (if valid) undermine our notions of moral and legal responsibility more broadly. That is, we can use evidence that is meant to establish that no one is free to reform our legal practice as a whole, e.g., by casting a critical eye on the retributive functions of the criminal justice system, but such general arguments are not applicable to individual cases aiming to exonerate a particular defendant. Neurosciences can, and they constantly do, give us further indirect, inductive evidence for physicalism. And physicalism can, in turn, lead us to challenge the ideas of agency and mental causation, and consequently the very idea of free will. Such a development could, ultimately, also have repercussions on law and legal practice.

Let us make a few clarifications before moving on. The following discussion will focus solely on the impact of neuroscientific evidence on assessing the level of legal responsibility of a defendant in criminal law. More precisely, the focus here is on the issue of the *culpability assessment* of an individual legal agent (natural person) in criminal cases. Although this is the most typical context in which the connection of law and the neurosciences is discussed, it is important to keep in mind that the issue is in fact much broader, and the neurosciences can affect legal practice in various different ways, and raise a number of different ethical and legal concerns (cf. e.g., Greely, 2009; Farahany, 2016; Greely and Farahany, 2019). Neuroscientific evidence can also be used in civil cases (e.g., as a part of benefit claims), and neuroscientific methods can be used, not only in assessing the defendant's mental state during the time of the criminal act, but also to improve our understanding of the behavior of other parties during court proceedings (i.e., witnesses, lawyers, judges, and juries), and to help us explain how the court arrives at its decisions (e.g., Schleim et al., 2011; Ginther et al., 2018). Neuroscientific evidence can also be used to inform our forward-looking judgments, e.g., in assigning punishment, in predicting and preventing criminal behavior, or in inducing neural changes (enhancement or impairment). Yet a different, but an important – and urgent – issue at the intersection of the neurosciences and law, is the question of how to regulate

the use and data management of various different computer-brain interface devices, and the issue of the relevance of artificial intelligence to the practice of law in general.

The following will also rely on a very broad understanding of the notion of "the neurosciences," encompassing e.g., anatomical, imaging (CT, EEG, fMRI, MEG, NIRS, PET, SPECT, X-ray), and behavioral considerations. "Neurosciences" will here also range across a variety of disciplines, from biology (phylogeny, ontogeny, physiology, genetics) to psychology, and the cognitive sciences in general. Although this does not depart from the general practice – as the discussion on the connection of law and the neurosciences typically relies on a very broad construal of "the neurosciences" – it is important to keep in mind that the field encompasses a wide range of methods and disciplines, and the distance between lower-level biological considerations and the higher-level psychological ones is significant. Indeed, the issue we are facing with respect to how to take neuroscientific considerations into account in our moral and legal reasoning can be seen to hinge on the very question of how our psyche should be understood to be related to its biological basis.

CONCEPTUAL PRELIMINARIES: FROM *ACTUS REUS* TO *MENS REA*

Intuitively, if one relies on a naturalistic view on the human mind, information about the neural basis of our decision-making and action-production should, in principle at least, have a bearing on our moral and legal reasoning. But why, exactly, would that be? What lies behind this intuition? Clearing up this conceptual landscape is the key to putting the empirical results and legal cases in their right places.

To zoom our focus, consider the following chain of conceptual dependencies:

> legal responsibility → moral responsibility → free will → agency → causation

Here is a way of unpacking these connections. For you to be held legally responsible, a harmful event must have occurred, and that event must have resulted from actions that you wilfully and freely decided to perform. That is, the right sort of causal connection must hold between your decisions to perform certain actions and the outcomes of those actions, and "[b]ecause moral responsibility is tied to such a natural relation (i.e., causation), and because the law is tied to morality, the law also is tied to this natural relation" (Moore, 2009, p. 5). That causal responsibility is entailed by both moral responsibility (e.g., Glannon, 1997, 2002; Sartorio, 2007, 2016; Driver, 2008a,b, 2012; Braham and van Hees, 2012; Szigeti, 2014; Whittle, 2018; Willemsen, 2019) and legal responsibility (e.g., Hart and Honoré, 1959; Feinberg, 1962; Moore, 1984, 2009; Fletcher, 1998; Lehmann and Gangemi, 2007; Simester, 2017) is not only widely shared assumption of moral philosophy and legal theory, but also constitutes a fundamental element of our moral psychology (e.g., Shultz and Schleifer, 1983; Darley and Shultz, 1990; Sloman et al., 2009; Malle et al., 2014; Lagnado and Gerstenberg, 2017; Willemsen, 2019).

It might be intuitively appealing to construe the connection between these notions wholly hierarchically, in terms of proper subsets (**Figure 1**). That is, one could think that for there to be agency (ability to act) there must be causal processes in the world (only some of which are agential), and for there to be free will, there must be agency in the world (only some of which is free), and for there to be moral desert, there must be freely willed actions (only some of which we bestow with moral desert), and, finally, for there to be actions that call for legal consideration, these must be deemed as morally reprehensible actions (only some of which are serious enough to call for legal action).[1]

Although some such hierarchy must roughly hold, one can also point to gaps. Consider, most notably, the connection between legal and moral responsibility: we are sometimes deemed legally responsible for harmful events that we are not deemed morally responsible for – at least not without important qualifications. We may, more precisely, be held *financially responsible* for the harm caused by our action – we may be required to compensate for the damage that has been incurred – even if we would not be held *morally blameworthy* for the action; we may be found liable in tort law, even if no crime has been committed. So, although the two notions are clearly intimately connected, moral and legal responsibility do not form a straightforward hierarchy.

What, then, separates these two types of responsibility from each other? Clearly: the agent's state of mind. More precisely: moral blameworthiness requires, not only that there is a causal connection between the agent's actions and the harmful outcome, but also that the agent strove purposefully (or at least negligently) to bring about the given outcome, and that she was aware of the harmful nature of the outcome. It is essential for moral blameworthiness, therefore, that a right sort of causal connection obtains between the agent's mental states and the outcomes of her actions. Legal responsibility can, in turn, take place in the absence of such a connection. Thus, it is useful, it is here suggested, to separate *liability* from *culpability* (**Figure 2** and **Box 1**).

Note that this distinction could in fact be stated even more starkly. One could hold that culpability (moral blameworthiness) actually has nothing to do with the actions of the agent and their outcomes – that it pertains solely to the agent's mental states, namely her desires and intentions, and the decisions she makes on the basis of them – and that liability (legal blameworthiness *simpliciter*), in contrast, has nothing to with the agent's mental states – that it pertains solely to the actions of the agent and the actual harm resulting from them (**Figure 2** and **Box 1**). On this construal, moral and legal responsibility would be completely

[1]To be clear, taking these notions to be connected in this hierarchical way does not entail that one is committed to some realistic or naturalistic way of interpreting the notions of moral and legal responsibility (cf. Moore, 1984). The thesis is merely that these notions are this way connected, whatever their ontological status. In particular, this does not commit one to rejecting legal positivism (and embracing the natural law view). One could hold that law (and morality – and causal explanation for that matter) are wholly socially dependent entities, but still maintain that they are connected in the manner outlined here. Note also that *any* view on law and morality must distinguish between them, and give an account of their connections. Indeed, it is one deeply entrenched misconception that legal positivism would be committed to the complete severing of the connection between law and morality (cf. Gardner, 2001).

FIGURE 1 | It is intuitive to think that legal responsibility is grounded in moral responsibility, which in turn requires agential responsibility, which is grounded in causal processes in the world. According to this simple hierarchical view, all the higher forms of responsibility are proper subsets of the lower level ones.

BOX 1 | The two aspects of legal responsibility.

according to the distinction introduced here, legal responsibility can be attributed to an agent either on the basis of liability, or on the basis of culpability, or both (**Figure 2**).

The necessary condition for an agent to be deemed liable (strictly liable) is a harmful outcome that has resulted from the actions of the agent; i.e., in order for an agent to be found liable, the actions of the agent must simply be causally connected to an outcome that is actually harmful (to another agent). In legal proceedings pertaining to liability, the status of the defendant is legal person (natural persons are a proper subset of legal persons). In case the defendant is found legally responsible in the sense of liability, she/it can be sentenced to compensate for the harm that resulted from her/its actions.

For an agent to be deemed culpable (morally blameworthy), in contrast, no harmful outcome need have resulted from the actions of the agent; inchoate crime is also held as a crime. Thus, the necessary – and, *prima facie*, also sufficient – condition for culpability is the mere *actus reus* (the guilty act), which in turn presupposes *mens rea* (the guilty mind; criminal intent, encompassing criminal negligence) of the agent. In legal proceedings pertaining to culpability, the status of the defendant is natural person (legal persons cannot be deemed culpable). In case the defendant is found legally responsible in the sense of culpability, she can be sentenced in accordance with the penal code (which can be understood to function in terms of retribution and/or deterrence and/or protection and/or rehabilitation).

In typical criminal cases both of these components of legal responsibility are present (for typically a harmful event has actually occurred), and guilty defendants are sentenced both to suffer penalties for their *acti rei* and to compensate for the harm that has resulted from them. However, these two components of legal responsibility are conceptually and historically distinct, and subject to different legal principles (which does not prevent them from seeping into each other in legal practice, e.g., in the notion of strict criminal liability, in assigning punitive damage compensation, or in the actual outcomes of *acti rei* having effect on sentencings in criminal proceedings).

separate notions. In fact, this is how it used to be in early Anglo-Saxon law, for example, where legal actions were only carried out to determine the level and the subject of compensation of a harm that had been incurred, and the way that the harmful event occurred, and the intentions of the defendant whose actions resulted in the harmful outcome were simply irrelevant for the proceedings (Walker, 1968; Jacobs, 1971).

Today, however, the situation is quite the opposite: assessing the level of culpability of the defendant plays a major role in criminal cases – the more severe the case, the more so. Establishing the motive of a crime, for example, can be crucial for finding the defendant guilty for the crime. That is, for finding the defendant guilty of a crime – in the sense of finding her culpable for it – the defendant must be found to have acted on the basis of the right sort of reasons during the time of committing the crime. Moreover, and more strikingly, also inchoate crime is held, in severe cases, as a crime. That is, our current legal systems can, contrary to the old Anglo-Saxon one, focus solely on assessing the level of culpability of the defendant – even to the point of ignoring the issue of whether a harmful outcome actually resulted from the defendant's actions. In fact, many claim that that is all that they should do, at least in criminal cases (e.g., Husak, 1987, 1998, 2007, 2011; Ashworth, 1993; Feinberg, 1995, 2003; Morse, 2004b; Alexander and Ferzan, 2009, 2018; Alexander, 2011; Enoch, 2014; Levy, 2015; Khoury, 2018).

The idea that we arrive at is the basic principle of criminal law: *actus reus non-facit reum nisi mens sit rea* – a harmful action without a guilty mind does not make one guilty. What this principle entails is that in order for a person to be found guilty of a crime, the right sort of causal connection must obtain (must be objectively shown to have obtained) between the defendant's mental states (desires, intentions, decisions) and the harmful outcome of the defendant's actions: *mens rea* – having a guilty mind – is necessary for culpability (**Figure 2** and **Box 1**). Consequently, a critical element of criminal proceedings often pertains to the issue of establishing the *criminal intent* of the defendant: to be found guilty of a criminal act, the defendant must have made a conscious decision to act in a way that would bring about the harmful outcome in question – and the outcome must have resulted from this conscious decision.

Another way of putting this idea – the central idea of the criminal justice system – is that the focus of the proceedings is on the question of whether the outcome under scrutiny happened due to a given agent: whether the harmful event occurred or not was under the agent's control. Central to the assessment of the level of culpability of an agent is, therefore, the notion of *sense of agency* – the issue of whether the occurrence of the harmful event was *up to the agent*. To be yet more precise, in order to be culpable, the mental states of the agent must have functioned as sources, or *difference-makers* for the outcome of

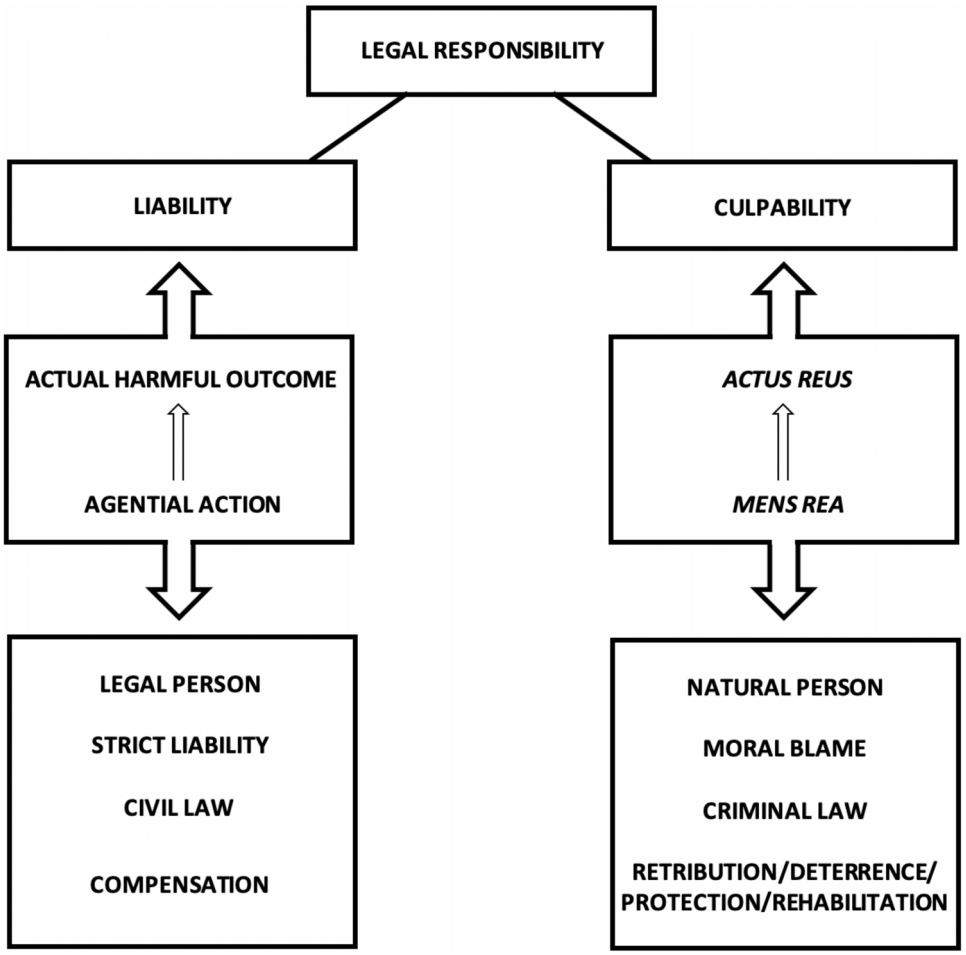


FIGURE 2 | Legal responsibility can be interpreted to have two elements, or aspects: culpability and liability (**Box 1**).

the agent's actions that have been deemed harmful. We can define "difference-making" in the following way: one event (the cause-event) is a difference-maker for another event (the effect-event) just in case the latter event would not have occurred had the former event failed to occur. For an agent to be culpable – for the right sort of causal connection to hold between the agent and the harmful outcome of her actions – the agent's mental states (desires, intentions, decisions) must have stood in a difference-making relation to the outcome: the outcome would not have occurred, this idea requires, had the agent been in a different mental state (and would therefore not have performed the action that led to the harmful outcome). Let us suppose that having this sort of a relation between the agent and the outcome of her actions is a minimal requirement for finding the agent culpable for her actions.

Now, given this setting, it is quite easy to see how neuroscientific considerations might start to have a bearing on legal reasoning: they help us to assess whether the right sort of causal relation obtained between the agent's mental states and the outcomes of her actions – they help us to assess whether the agent is culpable. Neural considerations might point to *neural dysfunctions* that could have disrupted the normal functioning of the neural basis of the relevant cognitive processes of the defendant. This could lead us to conclude that the required sense of agency did not take place – that whether the event deemed as harmful occurred or not was not up to the defendant – and the right sort of causal relation – the difference-making relation – between the agent's mental states and the outcomes of her actions was severed.

Although it should be clear that the notion of sense of agency is crucial here, this notion lends itself to different interpretations. Let us call one view on it *subjective* or *internal*, and another *objective* or *external*. In some cases of loss of agency, we are speaking in former terms: that the person did not feel, from her own perspective, as if she had been in charge of the given events. In other cases of loss of agency, we have the latter view in mind: that the person was not, regardless of how she felt, in charge of the given events. Both of these types of considerations can play a role in assigning agency, and both views can be relevant to culpability assessments. One could, however, make a case for holding the latter to be more fundamental. Consider, for example, the fact that schizophrenia patients often report having control

over things that they do not, in any objective or external sense, have control over (e.g., Voss et al., 2010). If you are under such a delusion, you are not (typically) considered to be culpable for the given harmful events. This would seem to suggest that external considerations can override reports about the subjective sense of agency, at least in assessing culpability: whether the occurrence of a harmful event was up to you is not, if you will, up to you.

The importance of separating these two different points of view on agency can be further demonstrated by considering how our moral and legal reasoning tackles intoxication. External considerations sometimes speak against the exemption of a defendant: even if the defendant's sense of agency (e.g., memory, self-control) would have been significantly impaired when acting under the influence of alcohol or drugs, in typical cases that would not lead to us to relieve her of her moral and legal responsibility. Why? Because the agent had control over inducing those states on herself. The situation changes completely, of course, if the agent had become intoxicated and had acted precisely the same way, but she had gotten to that state by being drugged, without her knowledge, by somebody else. This suggests that our responsibility attribution practices track the ultimate agential source of our actions and states of consciousness whence the actions flow (Dimock, 2012).

This leads to another, perhaps the most fundamental conceptual distinction, which ought to be kept sharply in mind: we must separate the notion of agency *simpliciter* from the notion of free agency. That is, it is one thing to establish that a person has agency, and another, further thing, to establish free agency: as the hierarchy outlined above suggests, only some forms of agency can be marked as "free." The notorious philosophical problem of free will pertains, first and foremost, to the latter notion: not many people are willing to strip us of agency – the ability to act – but many find it deeply problematic to attribute free agency – the ability to act freely – to us. What, then, is free agency? That is not an easy question, and no exhaustive answer to it will be given here. It should be noted, however, that both compatibilist and incompatibilist accounts, and various sorts of each, are out there (**Box 2**, **Figure 3** and **Table 1**). We remain neutral to this dispute – the issue of whether determinism is compatible with free will – and merely point out that both accounts must give some story about how free, responsible agency differs from agency *simpliciter*. It should also be clear that the sort of external considerations pointed to above are crucial here: both accounts agree that external manipulation and coercion – the right sort of external forces – can rob us of our freedom and affect our assessments of culpability.

To illustrate the importance of keeping these conceptual distinctions in mind in this context, consider the following example:

> "To be found guilty in the U.S. legal system, a defendant must not only have performed a prohibited act, she must also have done so in a legally culpable state of mind. For example, if Mary suffers an unexpected seizure while standing on a subway platform and bumps into John, causing him to tumble to his death beneath the wheels of an oncoming train, Mary is not guilty of murder. Yet if she purposefully gave the same bump to John, intending his death by subway car, she would be. Neuroscience has sometimes been taken to suggest that the two scenarios are fundamentally the same and that therefore the legal outcomes should also be the same. Here is the reasoning: the motives that led Mary to push John purposefully onto the train tracks are products of her brain, which was in turn shaped by her genes and her environment, neither of which she chose. Accordingly, she is no more 'responsible' for her act when she intends it than she is when she has an uncontrollable seizure" (Jones et al., 2013, p. 17628).

One can now propose the following conceptual breakdown of this example. If Mary intentionally pushes John onto the train tracks, fully aware that that would result in great harm to John, most likely his death, and this is the actual outcome of her actions, then we should find Mary culpable for her actions and criminally liable for their outcome – we should find her guilty of murder. If, in contrast, Mary suffers a seizure, or if her behavior is determined by some other force outside of her control – if she herself had been pushed by someone else, for example – then we should deem her lacking *mens rea*, and not find her culpable for he actions and liable for the harmful outcome that actually occurred due to them. The question now is: to what extent should we let neuroscientific considerations affect our judgment in placing Mary into these two contrary slots? Should we think that the neurosciences reveal that she is – or that all of us would be in similar circumstances – pushed by her brain (together with her genes expressed in the given environment) to act in a certain way, and should she therefore be exempted from culpability, no matter what her internal states of mind had been? Naturally, we are prone to answer in the negative. But in seeing why the answer should, at least typically, be no, we need to get a clearer sense of when, and why, something counts as an external cause of an agent's behavior and when we are merely giving an explanation of the way in which the behavior, and its psychological antecedents, are physically realized.

It seems clear that the key to unraveling all this is in pinning down the factors that lead us to strip a person of her agency. Being manipulated, or being physically pushed, by another agent will, in typical cases, make us conclude the person was not responsible for her actions and their outcomes. So, why, then, should biological factors sometimes be seen to play a similar role in stripping persons of their agency?

LESSONS FROM PHYSICALISTIC MONISM AND METHODOLOGICAL DUALISM

The fundamental problem with connecting neuroscientific evidence to psychological and behavioral data, and drawing conclusions about causal relationships between the two, is the following: we know that all our mental states and processes, our personalities, desires, beliefs, and decisions to act this or that way, are grounded in our brains. Who we are, and what we do, is wholly dependent on our brains – without our brains, we, and our

BOX 2 | Different accounts of free will.

Skepticism: accounts holding that we lack free will

The Hard Incompatibilist View

The sort of freedom required for moral responsibility is incompatible both with determinism and with indeterminism. So, however, the universe turns out to be, there can be no moral responsibility (Waller, 1990, 2011; Pereboom, 2001, 2014; Levy, 2008, 2011; Caruso, 2012, 2016, 2017, 2019; Shaw et al., 2019).

The "Willusionist" View

The sort of freedom required for moral responsibility is taken to be undermined by neuroscientific evidence, such as Libet experiments (Libet, 1985, 1994, 2002, 2003, 2004, 2006; Soon et al., 2008; Koenig-Robert and Pearson, 2019), which are taken to show that our conscious thoughts are not involved in producing our volitions (Wegner, 2002, 2004; Caruso, 2012).

Compatibilism: accounts holding that free will is compatible with determinism

The Hobbesian View

Freedom requires the ability to act on the basis of one's choices, free from external constraints and impediments (Hobbes, 1651/1994). An external constraint is a factor that prevents one from carrying out one's will; e.g., imprisonment might constrain an agent from acting as she wills. This view essentially rejects the notion of freedom of the will in favor of the notion of freedom of action; according to it free will is freedom to perform the actions we want to perform.

Conditional Leeway View

Popular view among a wide range of theorists especially in the first half of the 20th Century (Moore, 1903; Schlick, 1930; Ayer, 1954; Smart, 1961; Lewis, 1981; Berofsky, 2002). Freedom requires the ability to do otherwise, understood according to a conditional analysis of that ability. That is, an agent is able to act otherwise provided that the agent would have acted otherwise (or would be likely to have succeeded in acting otherwise) had she chosen to, or had she tried to.

Dispositional View

According to the dispositional analysis, freedom requires the ability to do otherwise, where this is analyzed in dispositional as well as conditional terms (e.g., Vihvelin, 2004, 2011, 2013). That is, an agent could have done otherwise if she would have done otherwise had she tried to, *and* if she could have tried to do otherwise. The ability to choose otherwise is then analyzed in dispositional terms: an agent could choose to do otherwise if she would choose otherwise were she placed in certain circumstances where the right sorts of triggers are present.

Hierarchical Control View

According to the hierarchical control view an agent's first-order desires (e.g., "I want a cigarette") must be distinguished from their second-order desires, desires regarding which first-order desires one has (e.g., "I want to not want a cigarette"). An agent's *will* is defined as the first-order desire that actually moves one to action. An agent has a second-order *volition* when that agent has a desire regarding which of her first-order desires moves her to action (i.e., has a preference about which of her desires becomes her will). On this view, an agent has free will insofar as she is moved by the desires she wants to be moved by; acting in accordance with free will is essentially acting on the basis of desires that one endorses (Frankfurt, 1971).

Real-Self View

According to the real-self view, it is not enough that one is moved by second order desires. What matters is that one's choices are in line with one's most fundamental system of values – the "real-self." These are the desires that one *rationally* identifies with (Watson, 1975).

The Reason-Responsiveness View

Fischer and Ravizza (1998) analyze moral responsibility in terms of "reasons responsiveness." That is, the ability to respond to reasons in such a way that one would have done what there is most reason to do even if circumstances had been slightly different. The account parallels Nozick's (1988) truth-tracking account of knowledge, according to which a belief counts as knowledge if it "tracks truth" in nearby possible worlds. Similarly, an agent counts as morally responsible (and hence having free will) if the agent's decision-making mechanism tracks reasons.

Emergent Freedom View

An emergentist view on free will concedes, in line with incompatibilism, that indeterminism at the level of agency is necessary for free will and moral responsibility. However, the view also holds that indeterminism at the level of agency is consistent with determinism at the lower levels of reality. This is possible, according to the view, because the agency-level phenomena are multiply realizable at the lower levels, and the same agency-level phenomena could therefore have been realized by various different underlying physical bases (List, 2014, 2019).

Asymmetric Accounts

The "Reason View"

According to the "Reason View" moral responsibility requires the ability to do the right thing for the right reason (Wolf, 1980, 1990). This principle is asymmetric in its compatibility with determinism, with respect to moral desert (praise or blame). If an agent *has* done the right thing for the right reason, then, *a fortiori*, she is *able* to do the right thing for the right reason, so the condition is automatically met in the case of praiseworthy action. In contrast, if the agent has done something wrong, then she has failed to do the right thing for the right reason. In this case, the agent will only be responsible if she was *able* to do the right thing for the right reason. This is read as requiring the ability to do otherwise, holding the past the laws constant. Hence, praise is compatible with determinism, but blame is not.

Incompatibilism: accounts holding that free will is incompatible with determinism

Event Causal Incompatibilism

Event causal incompatibilists typically endorse similar conditions of free will to standard compatibilist accounts (such as the capacity to respond to reasons and to act in line with one's deeper values). But they also require that one's choices are not determined. On this view, it matters that one's choices have the right sort of causal history (that they are sensitive to one's values and reasons), but this history would not leave room for free choice unless one's choice was also left open (where this is analyzed in non-conditional/non-dispositional terms). That is, free will requires the ability to do otherwise, as things actually stand, holding the past and the laws of nature constant (Nozick, 1981; Kane, 1996, 1999, 2011; Ekstrom, 2003; Franklin, 2011a,b, 2014, 2018; Lemos, 2018, 2020; see also Mele, 1995, 1996, 2006).

Agent Causal Incompatibilism

According to agent causal incompatibilism, freedom requires that the *agent* causes her own choices and actions, where this cannot be analyzed in event causal terms. On this view, the agent as a whole, rather than her mental states, must be one of the relata of causation, and she must figure as a direct cause of her choices and actions. The agent is a substance, an "unmoved mover," able to influence her choices without being bound by any prior causal influence. The falsity of determinism is required, either because agent causation must involve non-conditional alternative possibilities (leeway incompatibilism), or because exercising free will requires one to be the "ultimate source" of one's actions (source incompatibilism) (Reid, 1788/1969; Chisholm, 1964; Taylor, 1966; Clarke, 1993, 1996, 2000, 2019; O'Connor, 1995, 2002; Griffith, 2005, 2010; Steward, 2012).

Non-causal Views

According to non-causalists, free choices must be uncaused. That is, they must not be explicable in terms of the causal influence of prior events at all, and hence cannot be determined. Agent's choices would not occur at random though, as non-causalists would require that choices must be *rationally* explicable. That is, an agent's choice must made on the basis of reasons. It is denied, then, that reasons explanation is a species of causal explanation (*contra* Davidson, 1963). Reasons-explanations are taken to be *sui generis* (Ginet, 1990; McCann, 1998; Goetz, 2008).

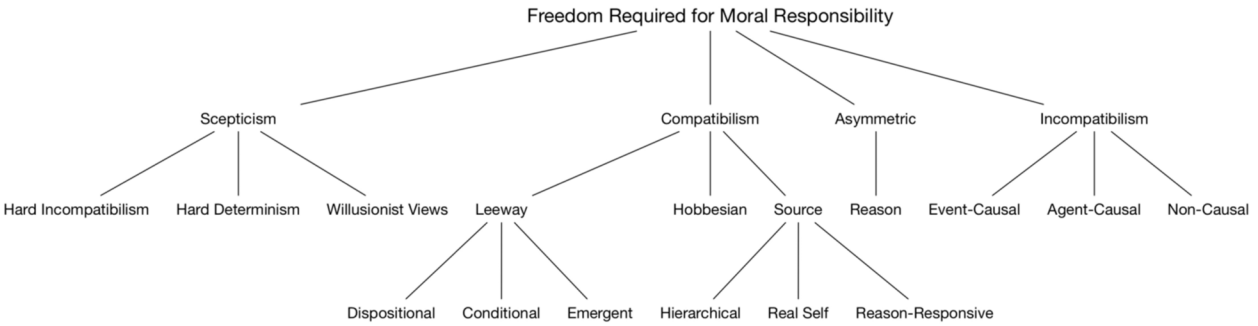


FIGURE 3 | A taxonomy of the different philosophical accounts of free will [adapted from Elzein and Pernu (2017)].

decisions and our actions, cease to exist. Therefore, just pointing to *some* neural correlates of our mental states or processes cannot, by itself, force us to conclude that those neural correlates, rather than the mental states that they ground, should be designated as the proper causes of our behavior.

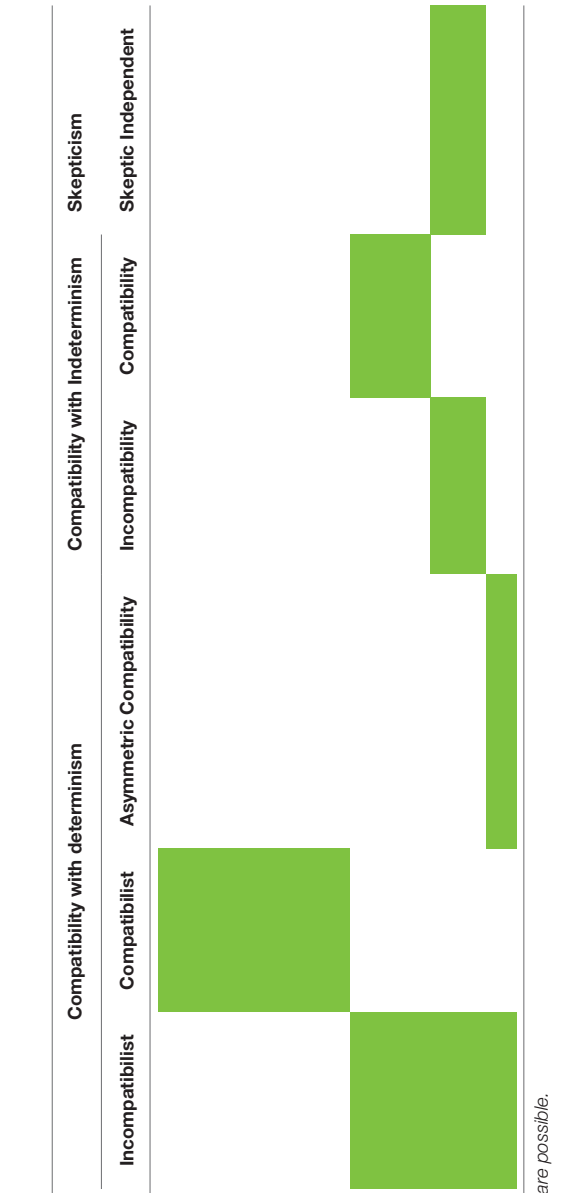
This point is rather trivial: if we assume physicalism, the view that we are biological, and ultimately physical entities, as it has been assumed here, then we can be sure that we will always find neural correlates for our psyche and behavior. And yet, the neuroscientific literature is rife with studies demonstrating the structural and functional differences of the brains of various different types of people, such as string players non-string players (Elbert et al., 1995), Braille readers and sighted (Sterr et al., 1998), taxi drivers and non-taxi drivers (Maguire et al., 2000), musicians and non-musicians (Gaser and Schlaug, 2003), jugglers and non-jugglers (Draganski et al., 2004), pedophiles and non-pedophiles (Cantor et al., 2007), hetero- and homosexuals (Ponseti et al., 2007; Savic and Lindström, 2008), adolescence-limited and life-course persistent antisocial behavior (Carlisi et al., 2020), murderers and non-murderers (Sajous-Turner et al., 2019), among others. It is often unclear what the import of these studies is. We can assume, as we already knew that these people are behaviorally homogeneous, that their brains, that ground their psyche and behavior, are in some respects homogeneous. Therefore, such an observation does not, by itself, support the idea that the behavior of these types of people is somehow essentially – more than in other, "normal" people – dependent on such neural factors.

Consider, to connect this issue to the topic at hand, the example of the vivid social and behavioral gender differences in criminology: it is well-known that men commit substantially more crime than women, across cultures (e.g., Rowe et al., 1995; Burton et al., 1998; Carrabine et al., 2004; Walker and Maddan, 2013). We also know that there are numerous significant physiological differences between the two sexes, including neural differences. Should we now conclude that men are more prone to crime than women, and, more importantly, should we maintain that it is the brains of men, rather than their conscious decisions, that make them commit these crimes, and that men are therefore less culpable than women for their criminal behavior, or maybe exempt from it altogether? This is not a generally accepted way of reasoning. But why not?

The question, of course, is this: how do the observed correlations between certain types of neural and mental states arise? There are two different, but connected issues here. First, there is the metaphysical issue of how we should understand the mind and its neural basis to be connected to each other. Second, there is the more pragmatic, or methodological issue of how we should determine the right order of causes and effects in this context. The first issue is more fundamental. Suppose that dualism is the right metaphysical view. Suppose, in other words, that the mental and the physical are wholly distinct from each other. Then the issue of how the neuroscientific (physical) evidence should bear on psyche and behavior would not arise at all: the mental realm would evolve according to its own laws (if any). Or suppose, in contrast, that the mental and the physical are identical. In this case both neuroscientific and psychological evidence would be completely translatable to each other (as they are assumed to be referring to one and the same thing).

Both dualism and the identity theory seem unacceptable to scientifically informed common-sense: neither are the mind and the body wholly distinct, nor are our mental notions completely translatable to neural ones (and *vice versa*) (cf. Pernu, 2017). But what could the third way be? According to non-reductive physicalism – arguably the received view in current philosophy – the mental is dependent on the physical, but non-reductively so. That is, according to this view, there is always some physical (neural) basis for mental states, but the mental cannot be reduced to, or identified with, its physical basis. What, more concretely, could this then mean? Typically, the connection between the two is supposed to be understood in terms of *realization*: the mind – its mental functioning – is realized by neural processes. The often-used analogy is the distinction between software and hardware in computation (e.g., Block, 1995): the mind is, close to literally, a software run by the hardware of the brain; the mind is what the brain does. It follows from this that although the mind, to be able to function, must always be realized in some physical way – like a computational software must be run by some hardware in order to be functional – it can be realized in different physical ways – like a computational software can be run by different types of hardware. So, mental states must be physically realized, but they can be *multiply realized* by a variety of physical states, and are not therefore reducible to, or identical with them. This is how, according to this view, we can both preserve the

TABLE 1 | Commitments of different accounts of free will.



*Compatibilists are typically event causalists though other positions are possible.

intuition that the mental and the physical are distinct, but avoid the conclusion that dualism, at least of the classical substance kind, must hold.

This non-reductive way of conceiving the relationship of the mind and its neural basis is certainly very attractive. It is not without problems, however. Not only has the thesis of multiple realization itself been challenged (e.g., Bechtel and Mundale, 1999; Shapiro, 2000; Polger and Shapiro, 2016), but the view has also been argued to be unstable precisely due to its reliance on multiple realization: it has been argued that in order to account for the causal efficacy of mental states, they cannot be conceived to be distinct from physical states, must be assumed to be ultimately reducible to them – on pain of deeming mental states wholly epiphenomenal (e.g., Kim, 1989a,b, 1990, 1993, 1998, 1999, 2005; Papineau, 1993, 2001, 2009). Consequently, a vehement debate over the status of non-reductive physicalism rages on in current metaphysics and philosophy of science (**Box 3**). This is not the place to declare a verdict on it. Nor do we need to: whether it is the reductive or the non-reductive sort of physicalism that will ultimately prevail, they are both committed to the thesis that is now of interest – namely the idea that mental states and processes are always, in one way or another, neurally realized.

Therefore, unless you are a card-carrying dualist, the simple fact that we can point to some neural correlates of our psyche and behavior should not come as a surprise. Yet, that is what often seems, implicitly or explicitly, to be the concrete conclusion of many brain imaging studies. Failing to keep clearly in mind the simple idea that mental states are always neurally based leads easily to the fallacious conclusion that it is discovered neural correlates that are causing the mental or behavioral differences that have been observed. But nothing of that sort can be established solely on the basis of the presented neuroscientific data. It is the biological function of the nervous system to be responsive to a variety of environmental cues, inanimate, animate, and social: it enables us to respond to the received stimuli in a flexible and appropriate manner. Different stimuli, together with a variety of preconditions at different levels of biological organization, shape our nervous system, which in turn forms the physical basis of our psyche and behavior. It is not so, therefore, that the changes in our psyche and behavior should be interpreted as being caused by neural changes, even if the two can be consistently linked.

To be more precise, in ignoring these basics, one easily falls prey to two different fallacies. First, there is the issue of the direction of causation. Once a neural correlate of a particular mental or behavioral feature is specified, one is easily led into thinking that the former is causally responsible for the latter – that the specified neural state *caused* the mental and behavioral changes that we observe. Typical neuroscientific data, imaging data in particular, is wholly statistical, and establishes only correlations between behavioral and neural variables, and the data alone does not therefore license a causal interpretation (cf. e.g., Tancredi and Brodie, 2007; Miller, 2010). Taking a more encompassing, metaphysical view on the issue does not give a shortcut to establishing causal conclusions. Although it follows from physicalism that mental and behavioral features are always

BOX 3 | Non-reductive physicalism and the problem of causal exclusion.

It seems natural for us to separate the mental from the physical, for various reasons (Pernu, 2017). For example, how you feel subjectively does not seem to be identical with the neural states that we observe to correlate with your feelings: although *you* might feel in a certain way – depressed, anxious, aggressive – it would not be correct to say that *your brain* has these feelings (even if these feelings would not be there without your brain). However, it also seems natural for us to hold that the mental and physical can interact – that your feelings and thoughts can have an effect on your body and on the course of events in the world surrounding you, and *vice versa*. But if we follow the first intuition, and set the two realms apart, it becomes difficult to see how they might interact.

How to resolve the tension between these two intuitions? Let us suppose – as both common sense and the current scientific consensus does – that neither the mental nor the physical can claim monopoly over reality. Let us suppose, in other words, that neither eliminative idealism nor eliminative materialism holds. If we assume that reality is neither purely mental nor purely physical, what options can we possibly have left? The currently popular view in philosophy suggests: *non-reductive physicalism*.

What, then, is non-reductive physicalism? Non-reductive physicalism holds that although the mental is dependent on the physical (in the sense that the former cannot exist without the latter), the former is neither identical with, nor reducible to the latter. But how can that be? How can something be dependent on something, but be neither identical with, nor reducible to it? Well, we can say that although no mental state can exist without being accompanied by a physical (neural) state, the reverse does not hold. That is, no particular physical (neural) state is necessary for a given mental state to exist. So, you cannot, according to this view, read off which particular neural state happens to hold from the psychological data alone, as a number of different neural states could function as physical bases of mental states. Although this view enjoys wide popularity in current philosophy of mind, it faces a well-known problem. According to the *causal exclusion argument* such a non-reductivist position is not stable, and will, when given a more detailed treatment, collapse into reductionistic physicalism (e.g., Kim, 1989a,b, 1990, 1993, 1998, 1999, 2005; Papineau, 1993, 2001, 2009). The source of this problem is in the basic assumption of all physicalism, namely in the assumption that the physical world is causally complete – that all physical effects have complete, sufficient physical causes. So, if every physical event in the world, that has a cause, has a physical cause that fully accounts for its occurrence, then postulating any mental causes appears wholly superfluous. It would thus seem inevitable that either mental states are epiphenomenal – that they are not doing any causal work in the world – or that they are identical with physical states – and as such states they would then be able to play the causal role we intuitively attribute to them.

The causal exclusion argument is currently under heavy debate. One popular non-reductionist strategy is to move the focus on the notion of causation at play in the argument, and criticize the idea that some events could be held "causally sufficient" for other events (e.g., Menzies, 2008, 2013, 2015; Woodward, 2008; List and Menzies, 2009; Raatikainen, 2010; Pernu, 2013b). If causation is understood in terms of counterfactual difference-making, rather than in terms of physical generation or production, the idea that mental states have autonomous causal power can be vindicated, according to this line of critique. However, there are a number of problems to address. There appears to be an equivocation on how the effect-events are individuated, for example, and the difference-making argumentation could be seen to lead to parallelism rather than interactionism (Pernu, 2013a, 2014a,b, 2016). And even more burningly, when the abstract philosophical argumentation is brought down to a concrete, neuroscientific level, the basic message of the causal exclusion argument appears to have bite again, and the mental and neural can be deemed identical, even if causation is understood in purely difference-making terms (Pernu, 2018).

The debate on how to relieve the tension between mental and physical causation, or higher and lower-level causation in general, continues.

neurally realized, and that the former are thus *constitutively dependent* on the latter, it would be wrong to think that the former must also be *causally dependent* on the latter – like, to use a very simple analogy, the bricks a house is made of are not a cause of the house. Note that even if one were to subscribe to the view that the mental reduces to the physical, or that the two are identical with each other, it would not follow that the former has to be causally dependent on the latter – quite the contrary: if the mental simply is the physical, then the two cannot be causally related, for identity is a symmetric relation (among other things) whereas causation is an asymmetric relation *par excellence*.

How, then, should we perceive the causal relationship between the two? That is not an easy question to answer. As already stressed, it might be altogether wrong to postulate any causal connection between the two (as the mental is realized, not caused by the physical). However, such a view would also go against the intuition that we often find it correct to say that the two are causally connected, e.g., when being knocked in the head causes you to become unconscious, or when being told that the house is on fire causes you to move your body out of the building. This is not the place to attempt to give a full account of how using such causal language – which is *prima facie* interactionist – can be made consistent with the monistic metaphysics of physicalism. It suffices here to make it clear that even in this physicalistic framework we need to give some such account: we need to explain why we sometimes point to physical (neural), and other times to psychological causes of our behavior. And more importantly: this precisely is the issue we are facing with neurolaw – the question of whether, in some cases, we should point to some biological *rather than* psychological sources of our behavior. That is, biological and psychological explanations of our behavior can, sometimes at least, be taken to be in a genuine pragmatic competition with each other.

This, it is here maintained, is at the heart of the problem of how to take neuroscientific considerations into account in our moral and legal reasoning. On the one hand, the discussion takes a certain kind of monistic metaphysics for granted, namely physicalism. On the other hand, we need to make sense of our talk of psychological vs. biological ways describing our behavior being mutually exclusive. Somehow, in other words, we need to accommodate our folk psychological dualism with metaphysical monism. Providing such an account is an ongoing philosophical project, and it is not the aim of this discussion to contribute to that. Here, we only point to this tension, and merely note that as long as we hold our folk psychological practices non-negotiable, and take our moral and legal reasoning to be resting on such practices, which seems plausible (cf. e.g., Lelling, 1993; Morse, 2003, 2004a, 2006, 2007, 2008, 2011a,b, 2013, 2015; Sifferd, 2006, 2018; Jakubiec and Janik, 2017; Hirstein et al., 2018; Moore, 2020), we must rely on *some*, albeit covert and unarticulated, criteria on how to demarcate between biological and psychological causal hypotheses. The monistic metaphysics of physicalism should therefore, in this context at least, be reconciled with methodological dualism.

It is, however, quite easy to point to one criterion that appears to play a crucial role in setting biological and psychological ways

of explaining behavior apart from each other. This is precisely the issue of sense of agency: if a lack of sense of agency is detected, one is prone to shift from the psychological realm to the physical realm in designating the source of the given behavior. If, in other words, some events are not under agential control, then their causal sources should be traced back to somewhere other than to the psychological factors. And, as already stressed, here external considerations bear considerable weight. Compulsion, coercion, manipulation, and other such chains of events where the ultimate sources of the outcomes we happen to be interested in are designated to lie out of the reach of the agent, rob that agent of agency – at least the sort of free agency that is central to culpability assessments.

This brings us to the other fallacy that an uncritical treatment of neuroscientific evidence easily leads to. As external considerations bear a significant weight in demarcating between biological and psychological causal hypotheses in explaining behavior, one might be led into thinking that pointing to a biological (neural) abnormality would count as clear evidence for the presence of a factor outside the control of the agent. That is, one easily makes an inference from consistent neural differences to the claim that those neural features must be causing the observed behavioral differences. However, no such conclusion can be made solely on the basis of abnormality considerations. This is precisely because the basic function of the nervous system is to react and adapt to environmental cues; psychological and behavioral differences will always manifest themselves as some neural differences.

There is, however, a connection here that is worth highlighting. The notion of abnormality (consistent patterns of difference) is closely related to the notion of *dysfunction*. Pointing to dysfunctional neural features – to dysfunctional biology – would seem to give grounds for concluding that it is these neural features, rather than the mental features of the agent, that we should consider to be the proper causes of the agent's behavior. Although this is no doubt the reason why abnormality considerations are prone to lead us to favor biological explanations over psychological ones, this observation will not take us much further in the analysis. The basic problem is that dysfunctional brains often reside in dysfunctional environments, and psyche and behavior can also be deemed dysfunctional. Again, therefore, pointing to a mere neural feature, a dysfunction in this case, cannot be made to justify the conclusion that this neural feature is the ultimate source of the behavioral features in question. We need independent reasons to hold the neural dysfunction to be caused by something outside the scope of the influence of the agent.

There is no doubt, however, that the notion of dysfunction is central here. Consider, in particular, the notions of disease and disorder, which are, according to a naturalistic reading at least, tied to the notion of dysfunction: a healthy organ or organism is one that functions properly, according to the way it's supposed to function in light of its ecological role and evolutionary history (e.g., Boorse, 1975, 1976, 1977, 1997). Diseases and disorders are in turn something that we quite naturally hold to be autonomous with respect to the psyche and behavior of the agent: they are something that happen to an agent, and hence they are not something that the agent is in any way responsible for. This, of course, is the reason why considerations related to mental illnesses and disorders are highly relevant to culpability assessments.

So, we can presume that in pointing to neural abnormalities to demarcate between biological and psychological causal hypotheses in explaining behavior, the chain of reasoning goes from abnormalities to dysfunctions, and from dysfunction to illness or disorder, and then from illness and disorder to an entity outside the scope of the influence of the agent. Now, although this might be taken to be the correct description of the actual reasoning that lies behind the tendency to infer from neural data – the sort of data that points to a neural abnormality – to the conclusion that favors the respective neural features over psychological ones, doubts can still be cast on whether this sort of inference should be endorsed. The problem is that it is perfectly legitimate to question the apparent value neutrality of the notions of health and disease (cf. e.g., Sober, 1980; Kingma, 2007). In fact, the very notions of function and dysfunction are notoriously difficult to define in thoroughly naturalistic terms (e.g., Mayr, 1988; Allen and Bekoff, 1995; Garson, 2016). The core of the problem is that in deeming an entity (a property or a process) either functional or dysfunctional, we are always relying on pitting *right* and *wrong*, or *good* and *bad* ways of performing the function against each other; there is a gap between the way the function is actually performing, and the way it is *supposed to* – the way it *ought to* – be performing. But this sort of a gap – the gap between ought and is – cannot be closed, as the history of philosophy teaches us (Hume, 1738; Moore, 1903). If this indeed is the case – if there is no way of finding a neutral, naturalistic basis for correct and incorrect ways of functioning – then it is an illusion to assume that our moral and legal reasoning could be based simply on identifying neural dysfunctions.

This issue is connected to a pragmatic, methodological problem that we face in attempts utilize neuroscientific data in moral and legal judgments. In assessing the neuroscientific data, we are engaged with the project of connecting such data to psychological and behavioral data. Although we are easily led into thinking that the former is somehow the more primitive and fundamental of the two – precisely because we are relying, tacitly of course, on a monistic physicalist metaphysics – it is in fact on the basis of the psychological and behavioral data that we draw conclusions about the function of the neural features that are being studied. That is, it is not so that we deem some neural features dysfunctional on the basis of the neural data alone – such data will typically demonstrate only that these features are statistically abnormal. We deem the features dysfunctional on the basis of our prior understanding of the psychological and behavioral features with which the neural features are correlated. We first deem, for example, psychopathy or pedophilia to be psychological and behavioral dysfunctions, and we then proceed to identify the neural correlates of such behavioral patterns, after which we deem those neural features dysfunctional – not the other way around. Given that this is the general pattern in which neuropsychological reasoning, imaging studies in particular, proceeds, it is very problematic to start basing our moral and legal judgments on neuroscientific data.

Not only are the psychological and behavioral considerations relevant, they are fundamental.

There is also an important trade-off to note: the more dysfunctional we consider the psychological and behavioral patterns to be, the less relevant the neuroscientific data related to these patterns is. If, for example, we are faced with psychologically and behaviorally clearly identified cases of mental illness or disorders (such as schizophrenia or psychosis), pointing to neural data correlated with these illnesses or disorders is bound to be thoroughly irrelevant to deeming the behavioral patterns in question as dysfunctional: we already know, based on the psychological and behavioral evidence, that the patterns are such. Consequently, in moral and legal reasoning – in making culpability assessments – related to such cases relying on the relevant psychological and behavioral evidence is wholly sufficient.

On the other hand, if we are facing less clear, or multifaceted psychological and behavioral patterns or personality traits (such as psychopathy or pedophilia), the neuroscientific data is bound to be irrelevant because that too is less clear and multifaceted. Morse (2011b, 2015) deems this the "clear cut" problem. To establish a reliable connection between behavioral and neural data, we need to rely on clearly defined behavioral variables. The less clear those variables are, the harder it will be to find robust and clearly defined neural correlates for them. But, when a sufficient clarity is achieved, and behavioral and neural data can be consistently connected, the neural data is bound to be irrelevant for our moral and legal reasoning with respect to the behavior – precisely because we already have a comprehensive and clear understanding of the behavior and deem it functional or dysfunctional wholly on its own merits.

Consider, finally, a concrete example of a case that exemplifies these problems we are faced with in trying to rely on neural evidence in our legal reasoning: the much-discussed case of *Roper v Simmons*, and the issue of whether we should hold the brains of adolescents underdeveloped, in relevant respects, and whether that should bear on our culpability assessments (**Box 4**). In this case, the defense tried to overturn a death-penalty sentencing ruling of a teenage defendant on the basis of arguing that adolescents have an impaired impulse control, due to their brains being underdeveloped, which should make us deem them less culpable than adults for criminal offenses.

The discussion on neurolaw often takes a critical view on the reasoning presented in the case (e.g., Glannon, 2011; Morse, 2011a, 2015). We can now see why that is: it exemplifies the very problems that were just reviewed in connecting neural data to psychological and behavioral data. The basic problem is, in other words, that we already knew that adolescents are different to adults in a number of ways, but in terms of impulse control, sensitivity to peer pressure, and sense of responsibility in particular. That is, we knew this based on our psychological, social and cultural understanding of adolescents, and as we knew that the psyche and behavior of us all, adolescents and adults alike, is dependent on our brains, it should not come as news to us that we can point to some neural differences that function as a physical basis of the psychological and behavioral differences that we observe. It seems that in *Roper v Simmons* the defense tried to make a case for causal explanation of the actions deemed harmful in biological, rather than in psychological terms, by appealing to neuroscientific evidence, in order to make the court infer a diminished sense of agency from this, and then make an inference to diminished culpability of the defendant

BOX 4 | Donald P. Roper, Superintendent, Potosi Correctional Center v Christopher Simmons.

Roper v Simmons [543 U.S. 551 (2005)] was a landmark ruling in which the Supreme Court of the United States held that it is unconstitutional to impose a death penalty for crimes committed by adolescents under the age of eighteen. The ruling was made when the defendant, 17-year-old Mr. Christopher Simmons, had appealed his sentence to be executed, after a jury had found him guilty of the murder of Mrs. Shirley Crook.

In the early morning hours of 9 September 1993, Simmons and his friend, 15-year-old Mr. Charles Benjamin, broke into Mrs. Crook's home, in Jefferson County, Missouri, as a part of a plan to commit burglary and murder. After Crook awoke upon hearing the pair and called out, Simmons and Benjamin entered her bedroom, tied her hands up, and covered her mouth and eyes with a duct tape. They then drove the victim to the Castlewood State Park, and pushed her off a railway bridge into the Meramec River, causing her death by drowning. They stole the victim's purse, which they later threw into the woods. The proceeds of the crime were reported to have added up to $6.

Both defendants were convicted for the crimes. Benjamin was sentenced to life in prison, but Simmons was given the death penalty. Simmons filed a series of appeals in the years that followed, and the case worked its way up both state and federal courts, with all of them upholding the death penalty. Eventually, in 2002, the Missouri Supreme Court stayed the execution while the U.S. Supreme Court decided *Atkins v Virginia* [536 U.S. 304 (2002)], which dealt with the issue of the death penalty for the intellectually disabled. As the U.S. Supreme Court did in fact rule that executing the intellectually disabled amounted to a cruel and unusual punishment, violating the 8th and 14th Amendments of the U.S. Constitution, the Missouri Supreme Court decided to reconsider Simmons' case, subsequently leading them to rule, 6-to-3, that executing minors would also amount to a cruel and unusual punishment. However, an earlier ruling of the U.S. Supreme Court, in *Stanford v Kentucky* [492 U.S. 361 (1989)], had decided that executing minors was not unconstitutional. This prompted the lawyers for Missouri, and Mr. Donald P. Roper, the superintendent of Simmons' correctional facility, to argue that the Missouri Supreme Court was contradicting the U.S. Supreme Court.

Therefore, in *Roper v Simmons* the question was, in effect, whether adolescent defendants should be considered analogous in relevant respects to the intellectually disabled in capital crime cases. Evidence was presented to the court aiming to establish that human brains, the prefrontal areas in particular, continue developing until the early twenties, and that minors are, for these precise neurodevelopmental reasons, biologically impaired in their capacity for moral reasoning and self-control. It was then argued that executing minors amounted to a cruel and unusual punishment, violating the constitution.

The U.S. Supreme Court did in fact overrule, 5-to-4, their earlier *Stanford v Kentucky* decision, and concluded that it is unconstitutional to impose a death penalty for crimes committed by minors, resulting in overturning death penalty statutes in 25 states. While neuroscientific evidence of the relative underdevelopment of the brains of adolescents was presented to the court, and while the arguments drawing on such evidence did receive significant attention, both from the experts and the public, the final verdict actually gave significantly more weight to social, psychological, and common-sense evidence, with the dissenting justices expressing skepticism of the relevance of neuroscientific evidence to legal procedure. However, the case demonstrated the potential impact of neuroscientific evidence to legal proceedings, and it was central in setting off the current discussion on the role of neuroscientific evidence in jurisprudence.

on the basis of this. However, such a chain of reasoning is not valid, precisely because simply pointing to neural differences between adolescents and adults should not make us conclude that a given action has biological, rather than psychological, causal etiology. Some independent reasons should be given to think that the biological features referred to are dysfunctional, and that this is due to biological rather than psychological factors. But as such reasons were not presented, all we are left with is the affirmation of the triviality that adolescents and adults have psychological and behavioral differences that are correlated with neural differences. Consequently, although the court did put some weight on the neuroscientific evidence presented by the defense (Carbone, 2011), its final decision was largely independent of it.

BASING LACK OF AGENTIAL CONTROL ON NEUROSCIENTIFIC DATA

The fundamental problem of utilizing neuroscientific evidence in our moral and legal reasoning stems from the fact that all decision-making and action-production is neurally based. It seems to follow from this that all our actions, including the *acti rei* that we find morally and legally concerning, are neurally caused. So, if simply pointing to such neural factors were to constitute a valid basis for exoneration, *all* our actions would become exonerable: "since all behavior is caused by our brains, wouldn't this mean all behavior could potentially be excused?" (Rosen, 2007). This is not how our actual moral and legal reasoning works. Typically, we are judged to be morally and legally responsible for our actions. But sometimes, in some cases, pointing to neural factors does have an effect on our moral and legal reasoning.

So, how to demarcate between good and bad ways of taking neuroscientific evidence into account in our moral and legal reasoning? Let us approach this question by considering some actual, concrete cases where it would seem natural for us to point to some neural factors as sources of behavioral patterns that we find morally and legally concerning (**Boxes 5**–**7**).

Consider, first, the historically important, and much discussed case of Mr. Phineas Gage, whose personality changed dramatically after a serious brain injury due to an accident in 1848 (**Box 5**). According to the sources of the time (Harlow, 1848, 1868), the once a hard-working, responsible, and much-liked man became, after the accident, explicitly anti-social and could not return to his previous job. His personality changed completely; "Gage was not," his friends would say, "Gage anymore."

Consider, next, the case of Mr. Charles Whitman, who indiscriminately shot at victims on a campus of The University of Texas at Austin in August 1966 (**Box 6**). Before the killings, he documented having "irrational thoughts," and feeling that he does not "understand himself." He requested that an autopsy be performed after his death to determine the cause of his thoughts and feelings, and his uncontrollable urge for aggressive behavior. A brain tumor was in fact later found, and it is plausible to suppose that Whitman may have suffered diminished control due to the tumor.

Consider, finally, the case of a man described by Burns and Swerdlow (2003), who developed uncontrollable and uncharacteristic sexual urges, that included pedophilic tendencies (**Box 7**). This led into him being arrested and convicted. Later, a brain tumor was found, and removed, which resulted in the disappearance of his criminal behavior.

One important thing to note is that anatomically all these cases involve some neural changes located in the prefrontal cortex (PFC). The function of PFC, in turn, has been associated emotional regulation and social behavior. Increase in aggressive behavior has been linked to PFC damage in Vietnam War veterans (Grafman et al., 1996), and reduction in PFC brain volume has been reported in patients diagnosed with anti-social personality disorder (Raine et al., 2000), aggression disorder (Woermann et al., 2000), and pathological liars (Yang et al., 2005). Imaging studies have revealed abnormalities in PFC function in violent people (Volkow and Tancredi, 1987; Chester et al., 2017) and convicted criminals (Raine et al., 1994). We have good reasons to believe, therefore, that changes in PFC are linked to anti-social and aggressive behavior (Brower and Price, 2001; Sapolsky, 2004; Hirstein et al., 2018). Interestingly, however, Ellenbogen et al. (2005) have reported a case of PFC lesion which resulted in a reverse change, namely a previously anti-social and violent individual turning into a docile and cheerful person (**Box 8**).

What this evidence suggests, to be precise – and all that it, by itself, suggests – is that there is a connection between the functioning of PFC and social behavior and aggression. Even if there were systematic differences in PFC in people deemed particularly anti-social and aggressive, compared to behaviorally and psychologically normal population, this should not, by itself, lead us to conclude that these people display the anti-social and aggressive behavior due to the changes in PFC, rather than the other way around (cf. Kishiyama et al., 2009). Moreover, the case

BOX 5 | The case of Mr. Phineas P. Gage.

Perhaps the most famous historical case demonstrating a dramatic change in personality and agential control is the case of Mr. Phineas Gage [1823 (presumed) – 1860], a 25-year-old railroad worker, who, in 13 September 1848, endured a devastating accident when an iron rod blasted through his head (Harlow, 1848, 1868; Macmillan, 2008). The rod entered through the left side of Gage's face, breaking his upper jaw, pushing directly through his forehead (destroying his left ventromedial frontal cortex), and protruding out through the top of his skull (**Figure 4**).

Astoundingly, Gage survived the incident. However, the physicians treating him chronicled dramatic personality changes, including a lack of restraint, and a marked decrease in his ability to control his impulses.

While this case has become legendary in psychology literature, it has also been apparently subject to notable embellishments. Nonetheless, the case still seems to provide a clear example of changes to behavior and capacity for self-control that likely result from brain injury (Damasio et al., 1994; Macmillan, 2000, 2008). Phineas Gage's skull is now on display at the Warren Anatomical Museum, Harvard Medical School (**Figure 4**).

BOX 6 | The case of Mr. Charles Whitman.

Mr. Charles Whitman (1941–1966) was a student at The University of Texas, with a previous career in the Marine Corps. He was largely described as a popular and intelligent young man by those close to him.

On the night of 31 July 1966, Whitman drove to his mother's house and stabbed her to death. He then went back home, and stabbed his own wife to death. That night Whitman typed notes in which he proclaimed to love his mother and his wife very much, despite brutally killing them both. He also expressed his inability to understand or explain his own behavior, and requested that an autopsy be performed in order to determine whether there was some biological cause for his actions, which might also explain the constant headaches he had been suffering.

Next morning, 1 August 1966, Whitman, a skilled marksman, climbed to the 28th floor of the tower of the main building at The University of Texas at Austin, and began shooting indiscriminately. He ended up killing fourteen people and injuring a further 31, before being killed by a police officer.

In an autopsy conducted after his death it was discovered that he had had a brain tumor. This was classed as a glioblastoma multiforme tumor the size of a pecan, located beneath the thalamus, but potentially impacting the hypothalamus, the temporal lobe, and the amygdaloid nucleus. Many have dismissed the tumor as being unlikely to explain his behavior, in line with the original conclusion of Dr. Coleman de Chenar, who first performed the autopsy. Nonetheless, Texas Governor John Connally's committee, comprising thirty-two experts, argued that the case is inconclusive (Texas Governor's Committee and Consultants, 1966).

Those doubtful about the significance of the tumor point to a number of psychosocial factors, such as Whitman's troubled relationship with his father, his anger at his life situation, feelings of personal failure, and his domineering behavior toward his wife (Lavergne, 1997). But such explanations do not obviously help to account for his explicit claims not to understand his own behavior, his explicit record of struggling to control impulses he failed to recognize as his own (as chronicled in his diaries and suicide note), and the fact that he was actively trying to seek psychiatric help for his condition. There is also a great deal of neural evidence that does in fact link disruptions to the amygdala and temporal lobe to aggressive behavior, rage, and poor impulse control (e.g., Damasio, 1994; Grafman et al., 1996; Anderson et al., 1999; Bechara et al., 2000; van Elst et al., 2000; Mobbs et al., 2007; Schneider and Koenings, 2017).

BOX 7 | The case of recurring brain tumor and pedophilia.

Burns and Swerdlow (2003) describe the case of a 40-year-old male who began to develop a strong interest in pornography, including child and adolescent pornography, and began to make sexual advances toward his prepubescent stepdaughter (leading to a conviction for child molestation). The man had no previous record of sexual interest in children. His behavior was coupled with a broader inability to control his sexual impulses, and with attempts to solicit sexual contact in inappropriate circumstances.

The person was eventually admitted to hospital, on the basis of complaining about a headache. While at the hospital, he reported balance problems, displayed marked difficulties with some of his movements, and appeared unconcerned that he had urinated on himself. He also had suicidal thoughts, reported fearing that he would rape his landlady, and attempted to solicit sexual favors from female nursing team members.

Magnetic resonance imaging (MRI) scans revealed a tumor displacing his right orbitofrontal cortex and distorting the dorsolateral prefrontal cortex. Upon removal of the tumor, his bodily control returned to normal, and after participating successfully in a Sexaholics Anonymous program, he was believed to pose no more threat to his stepdaughter, and was able to return home.

A year later, he again developed a consistent headache and began secretly collecting pornography. MRI imaging showed tumor regrowth, and once again his symptoms abated after its removal.

BOX 8 | The case of suicide attempt with a crossbow.

Ellenbogen et al. (2005) describe a case of a suicide attempt with a crossbow. Although the victim, a male in his early 30s, survived, he suffered a severe brain injury, as he shot himself to the head through his lower jaw. The bolt penetrated through the front of the victim's head, but did not exit through the top of the skull.

The result was a prefrontal cortex (PFC) injury, which gave rise to a dramatic personality change. The victim had a record of violent and anti-social behavior. After the injury, however, his behavior changed to the opposite: he became docile, social, and "inappropriately cheerful." This is in stark contrast to the typically described cases where a lesion to PFC results in aggression and anti-social behavior (**Boxes 5**–**7**).

described by Ellenbogen et al. (2005) indicates that similar types of damages to PFC can actually manifest in completely opposite psychological and behavioral changes (**Box 8**).

There is, however, one notable issue connecting the cases described above (**Boxes 5**–**7**). These are cases where the normal – a particular person's previous – functioning of the PFC has become disrupted due to an injury (lesion) or a tumor. This is the central reason, it is here suggested, why we can take these sorts of cases to have an impact on our moral and legal reasoning. That is, due to the etiology of these conditions, we do not consider these as cases of "brain rewiring," and we point to unequivocally biological, physical causes for these conditions. Now, of course, the interesting question is: why do we feel justified in reaching such a conclusion? Several factors are bound to play a role here. For one, lesions and tumors are easy to localize; they are concrete, spatially extended, material entities – something paradigmatically non-mental. They are not fuzzy, and they do not come in degrees, in contrast to the corresponding psychological or behavioral features: nobody has a lesion or tumor "more or less," but people can be more or less anti-social, or be more or less good at exercising self-control. Considerations related to gradedness in psychological and behavioral features and their neural correlates often affect our moral and legal reasoning by making clear-cut judgments difficult (e.g., Glannon, 2011). In cases of lesions and tumors, however, we can point to precise differences, not only spatially, but also temporally: the observed psychological and behavioral changes are dramatic and sudden, and it is therefore natural for us to tie these changes together with the clearly localized neural changes. These are the reasons, at least some of the main ones, why in cases like these we are prone to point to physical, rather than mental, causes of these conditions.

But we can dig deeper. It is not just that in these cases we feel it is natural to see these conditions as stemming from physical, rather than mental causes, but a radical lack of agential ownership is also associated with the conditions and their causes. That is, it seems clear that one fundamental reason why we find cases like these relevant to our moral and legal reasoning lies in our intuitive feeling that the ultimate source of these

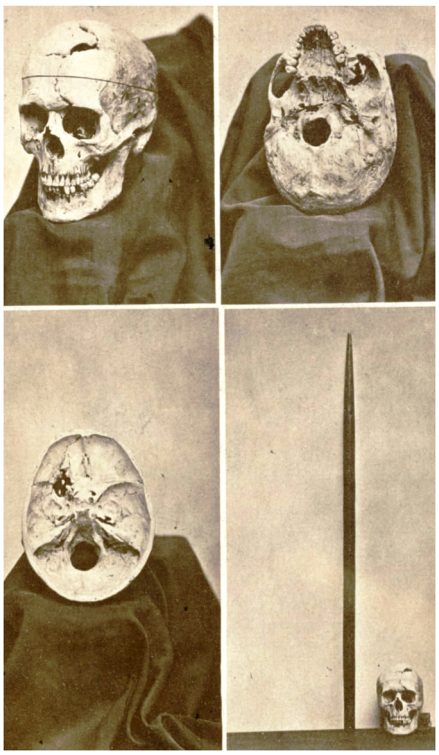


FIGURE 4 | The skull of Phineas Gage and the iron bar that pierced his skull in the accident on 13 September 1848, as shown in a catalogue of the Warren Anatomical Museum, Harvard Medical School (Jackson, 1870).

conditions must be placed outside the sphere of the influence of the agent in question. But why exactly that is, is not an easy question to answer. In the case of Mr. Whitman, for example, the internal sense of agency seems to have been lost (**Box 6**). But on the other hand, in the case of Mr. Gage (**Box 5**), and the case described by Burns and Swerdlow (2003) (**Box 7**), it is the external sense of agency, or the continuation of personality, that got disrupted. However, in both cases a sudden misalignment of feelings, thoughts, desires, values and actions occurred. This in turn affects the grounds of attributing agency to these subjects, and leads us to place the sources of their conditions outside the scope of their influence.

There is further important element to this way of reasoning. The described cases are cases of lesions and tumors. Lesions and tumors can be considered to be paradigmatic cases of dysfunction and illnesses. That is, in cases of injuries and diseases – such as brain cancer – we are automatically prone to think that something has gone biologically, rather than psychologically, wrong, and we exclude outright the possibility of neural rewiring. These conditions are neurological rather than psychological or psychiatric, and the proper way to intervene on them is physical (surgical, pharmacological) rather than psychological or behavioral. Although this is clearly an important issue affecting our causal and moral reasoning with respect to these cases, one can also envision caveats. The personality change described by Ellenbogen et al. (2005), for example, was due to a self-inflicted brain injury (**Box 8**). If it were – or when it becomes – possible to change one's personality traits by direct neural interventions, our attention is bound to shift more from the neural changes to the variety of ways these changes can be induced when taking account of neuroscientific evidence in our moral and legal reasoning [this would parallel the case of "grand schemers" – people who get themselves intoxicated before committing a crime in order to appear less culpable for it (Dimock, 2012)].

There are, therefore, a number of intertwined reasons why in the described cases we find it plausible to let neuroscientific evidence affect our moral and legal reasoning. One could argue, however, that in the case described by Burns and Swerdlow (2003) all the relevant issues come into play in the clearest possible manner (**Box 7**). Note, also, that this is the only case from the three where the neuroscientific considerations had a real, and significant effect on the proceedings (the case of Mr. Gage has no criminal component to it, and the case of Mr. Whitman never went to trial). In this case, the defendant was in fact acquitted on the basis of the presented neuroscientific evidence (after neuro-surgical interventions had been conducted). Why is this case special? We can point to two reasons. First, it involved a brain tumor, and there are a number of reasons why that has a bearing on our moral and legal reasoning, as just discussed. Second, and more importantly, the tumor could be designated to be the proper difference-making cause of the actions the defendant was accused of: not only did the behavioral patterns considered harmful disappear upon the removal of the tumor, they actually reappeared upon the reappearance of the tumor. This leads us to point unequivocally to the tumor, rather than the defendant, as the source of the actions he was accused of. And, as we hold tumors biological, non-mental entities, paradigmatically dysfunctional in the context of the biology of the person, we place the cause of the actions outside of the scope of the influence of the agent.

PHYSICALISM, FREE WILL, AND MORAL RESPONSIBILITY

The preceding analysis has been based on the pragmatic assumption of methodological dualism, the idea that it makes sense, in this context, to divide causal explanations into two groups, mental and physical (neural), and, at least sometimes, to point to one of them as the proper cause of behavioral patterns at the expense of the other. This is what we are poised to do when we cite neural changes as the basis of acquittance, as in the case described by Burns and Swerdlow (2003) (**Box 7**). Note, that in this case [in contrast to the case of Mr. Whitman (**Box 6**)], the defendant seems to have been motivated in performing the *acti rei*, and the diminished sense of agency was attributed to him on largely on external grounds (although he also stated having attempted to restrain his urges). One could, therefore, argue that the tumor was in fact part of him – his personality – and that it is wrong to see the issue in dualistic terms. That is, one could argue that the appearance of the tumor resulted in physical (neural) changes, which manifested as the personality changes, but that it is wrong to see these two different ways of describing the process

as distinct and in competition with each other. However, this is *not* how we, in practice, think. We consider the tumor to be alien to him, creating a biological, and, consequently, a psychological and behavioral dysfunction that calls for correction by physical means (i.e., surgery). So, even though the tumor was part of him – and an essential part of the physical basis of his personality – we take it to be a separate physical factor, and something that deserves, rather than the defendant as a person, to be designated as the cause of the actions deemed harmful.

But there is a deep conceptual problem with such an approach, as has already been stressed: such methodological dualism goes against the metaphysical monism of physicalism. It seems, therefore, that it is not ultimately tenable to hold that we can point to *either* mental *or* physical causes to our actions. Or, more precisely, physicalism seems to make it impossible to hold that there would be unequivocally mental causes to our actions: according to physicalism, all such causes must be physically realized. So, either mental causes – the psychological features we hold causally efficacious – must be reducible to, or identical with, physical causes, or they are not genuine causes at all, and we should altogether refrain from applying causal terminology to the psychological realm. On pain of eliminating our folk psychological practices, on which our moral and legal reasoning rests, we must, therefore, hold all our talk of mental causes to be covert talk of physical causes. But in that case the distinction between the good and the bad ways of applying neuroscientific evidence to moral and legal reasoning, as outlined above, would seem to collapse: in cases where pointing to psychological, social and behavioral factors, rather than to neural factors, as causes of our actions seems to us justified, it will not in fact be so, as we are dealing with physical causal processes through-and-through, and can therefore always point to physical causes of our actions. But if *acti rei* were always performed due to physical causes, and if pointing to such causes function as a basis for exoneration, then all *acti rei* should become exonerable.

This sort of reasoning can be seen to lie behind the more global worries related to the relationship of the neurosciences and jurisprudence (e.g., Greene and Cohen, 2004; Sapolsky, 2004). It is important, however, to make a clear distinction between two different ways of arguing from neuroscientific evidence to conclusions concerning moral and legal responsibility. The global worries are intended to prompt us to entertain doubts about free will and moral responsibility across the board. The argument would start with the assumption – typically a tacit one – that pointing to *any* physical basis for our psyche and behavior should make us cast doubt on moral and legal responsibility, and would then make the further assumption that relying on neuroscientific evidence constitutes such pointing by establishing physicalism. It is clear, however, that this type of argumentation would not make sense as a defense in any individual case, where the goal would be to seek grounds for exonerating the defendant on the basis that she does not have the capacity to exercise control over her actions in the way that is typically taken to be necessary for legal responsibility. In cases of this sort, the evidence only bears on the case insofar as it shows that the defendant is *abnormal* in contrast to typical defendants. But, as has been stressed, merely pointing to neural correlates, by itself, tells us very little about the causal source of the agent's actions, and on its own provides no grounds for assuming that the factors in question are outside the scope of the agent's influence. Insofar as global worries are to be taken seriously, then, they would need to be backed up with a much more speculative argument: one aiming to establish that even ordinary neural functioning ought to be regarded as inconsistent with moral and legal responsibility.

Global worries related to the relationship between the neurosciences and jurisprudence are, therefore, unlikely to be of practical importance, at least in individual cases. This explains, partly, why neuroscientific evidence has had much less bearing on actual legal practice than one would maybe have expected. It is worth outlining, however, in a bit more detail, the sort of reasoning that could be seen to give rise to such worries. Note, firstly, that although the discussion here has simply assumed that physicalism holds – that is, it has been assumed that everything is ultimately physical – it is not at all a trivial philosophical project to try to pin down where such an assumption stems from. And indeed, one could argue that physicalism is an empirical thesis, albeit rather holistically and indirectly such, and that the development of the neurosciences, in particular, has played a crucial part in making us reject dualism and persuading us to believe in the monistic metaphysics of physicalism instead (cf. e.g., Papineau, 2001, 2002). It is plausible, therefore, to connect empirical evidence, and the results of the neurosciences in particular, to establishing physicalism. But note that building such a connection is an incremental process, and although some particular results could be seen to bear more significance to it than others – such as connecting electric stimulation to muscle contraction (Galvani, 1791, 1794), or identifying neural cells as the units of the nervous system (Ramón y Cajal, 1888; López-Muñoza et al., 2006), or inventing neuroprosthetic devices (Shenoy et al., 2003; Hatsopoulos et al., 2004; Musallam et al., 2004; Pernu, 2018) – the thesis is not being proved, or disproved, by any single piece of empiria.

However, even if one thinks that there is such a connection between the empirical results of the neurosciences and the metaphysical thesis of physicalism, it is much more contentious to claim that physicalism, by itself, would disprove our ideas of free will and moral responsibility. That might be the case, but it would need to be argued for much more thoroughly and precisely, and it is definitely not a position that would enjoy wide-acceptance in the current discussion – even mental-to-physical reductionism is often motivated by the aim of saving the ideas of mental causation and agency (e.g., Kim, 2005, 2007; **Box 3**).

Most importantly, however, there is a metaphysical equivocation in here, that the discussion tends to overlook: mental-to-physical reductionism is not psychology-to-neuroscience reductionism. That is, even if we were to subscribe to a thoroughly physicalist metaphysics, we would not be committed to the idea that by inspecting the brains – or even the whole bodies – of people, we can in an any meaningful way, let alone perfectly, read their psyche. What we are facing here is the very same problem we have been facing all along: our brains, and our bodies, are built to react to environmental cues. So, whether a bodily state – a physical state – represents something meaningful, is not something that hinges on that bodily state

BOX 9 | The people of the State of Illinois v Nathan F. Leopold Jr. and Richard Loeb 33623/33624.

One of the most infamous cases in criminal history (e.g., Higdon, 1975; Theodore, 2007) occurred in Chicago in May 1924, when 19-year-old Nathan Leopold (1904–1971) and 18-year-old Richard Loeb (1905–1936) conspired in the kidnapping and murder of Robert "Bobby" Franks (1909–1924), a 14-year-old neighbor and second cousin of Loeb.

Leopold and Loeb were students at The University of Chicago, both wealthy and high academic achievers, with Leopold often described as a child prodigy, and with Loeb skipping ahead many years in school and becoming the youngest graduate of the University of Michigan at the age of seventeen. They spent several months planning the kidnapping and murder of their victim, and they were determined to commit "the perfect crime," simply for the thrill of it. They were inspired by the works of Friedrich Nietzsche (1844–1900), with Leopold supposing that their superior intelligence meant that they were "Übermenschen" and above the social and moral conventions that bound average, unexceptional people. Despite their efforts to make sure they would not get caught, the police quickly found leads that pointed to the boys' guilt, and they soon confessed to the crime.

The trial at the Chicago's Cook County Courthouse courted sensational media coverage during the summer of 1924. The defendants hired a renowned defense attorney, Clarence Darrow (1857–1938), who was an outspoken opponent of capital punishment. He entered a guilty plea, but proceeded to persuade the judge to avoid sentencing the defendants to death.

The court proceedings of the case are interesting in two respects. First, this is one of the first criminal cases where psychological and neuroscientific evidence was presented in a trial (Weiss, 2011) – albeit some of it in a now debunked form of phrenology (**Figure 5**). Second, the concluding speech of the defense, presented by Darrow, is famous for its sentiment, rhetoric, and appeals to global worries about free will. The closing argument, which lasted for twelve hours, built such an emotionally strong case for the defendants that it left the judge himself in tears.

Argumentatively, the speech was based on Darrow's conviction that none of us are really the sources of our choices, but they, and all the actions we base on our choices, are rather fully determined by psychological, physical, and environmental factors outside the scope of our influence (Darrow, 1922). In the trial, Darrow therefore pleaded that the boys ought to be spared on grounds that focused primarily on societal issues, making relatively little reference to the unique circumstances of the defendants in committing the murder [it is notable, however, that according to some of the expert witnesses of the defense the defendants were emotionally impaired, and Darrow would later argue that proper emotional functioning is necessary for making well-founded choices (Darrow, 1932) – echoing some of the developments that have been taking place in the discussion in the last 20 years or so]. E.g., the speech drew on the claim that, in the aftermath of the First World War, society had increasingly glorified war, sending a message to young people that life is cheap and killing is trivial. The core of Darrow's argument is neatly summarized in the following passage from the speech:

> "Why did they kill little Bobby Franks? Not for money, not for spite; not for hate. They killed him as they might kill a spider or a fly, for the experience. They killed him because they were made that way. Because somewhere in the infinite processes that go to the making up of the boy or the man, something slipped, and those unfortunate lads sit here hated, despised, outcasts, with the community shouting for their blood" (Darrow, 1924, p. 22).

The speech was successful: instead of death by hanging, Leopold and Loeb were sentenced for a life in prison plus 99 years.

Loeb was killed in prison by a fellow inmate in 1936. Leopold was eventually released in 1958, and he completed a master's degree at the University of Puerto Rico, after which he worked in various teaching posts and research projects – even publishing a book on ornithology (Leopold, 1963). He died in Puerto Rico from natural causes in 1971.

alone. The context matters; the environment in which the body resides – and has resided – needs to be taken into account. Mental dysfunctions, in particular, are highly sensitive to various environmental factors. The problem is not just that the mind can be multiply realized by various different bodily states – that you cannot read the bodily state from the mental state – but that in fact the opposite holds too: that the same bodily state can realize various different mental states – due, simply, to neural plasticity and reuse – and you cannot, therefore, identify the mental state by simply nailing down the exact bodily state that happens to realize it [as illustrated by the varied psychological and behavioral changes resulting from PFC lesions, aggressive and anti-social cases (**Boxes 5**–**7**) in contrast to docile and social cases (**Box 8**)]. The surroundings of the body – and not just the current stimuli it receives, but all the environmental cues that it has historically been exposed to – play a crucial role in shaping the body. Neither the bodies, nor the brains as their proper parts, can, therefore, play the role of a proper, complete physical realizers of mental states.

Consider, to make this argumentation more concrete, the famous case of *Illinois v Leopold and Loeb* (**Box 9**). The case is known – apart from the morbidity of the crime – precisely for the issue of global worries related to free will being presented to, and having an effect on, the court. It is notable, however, that even though the case is also important for it being one of the first examples of biological and neuroscientific evidence being presented to the court as a basis of the culpability assessment of the defendants (Weiss, 2011; Wilson, 2015), this did not, once again, have an effect on the final decision of the court. What did play a part in the ruling, instead, were more general environmental and social considerations, related to the age of the defendants in particular. In this way the case is actually quite strongly analogous to the more recent case of *Roper v Simmons* (**Box 4**).

None of this should make one conclude that physicalism does not hold. It should only prompt one to reject the idea that mind-body reductionism holds. To have a complete, firm grip on the mind, being in possession of a complete physical description of the body is not enough. You also need to be in possession of the complete description of the body's surroundings, and the history of the interactions of the two. The mind – its content – in other words, is, in physical terms, not only dependent on the nervous system that happens to realize it, but also on the environment to which that nervous system has been adapted. All this can, in principle, be described in physical terms, as both the body and its surroundings are, in the final analysis, physical entities. In practice, however, the interactions are too intricate, and the system as a whole is too complex, for us to be able to make sense of it in purely physical terms.

The mental might, therefore, be reducible to the physical, but it won't be reducible to mere bodily states. That is why the global worries related to the relationship between the

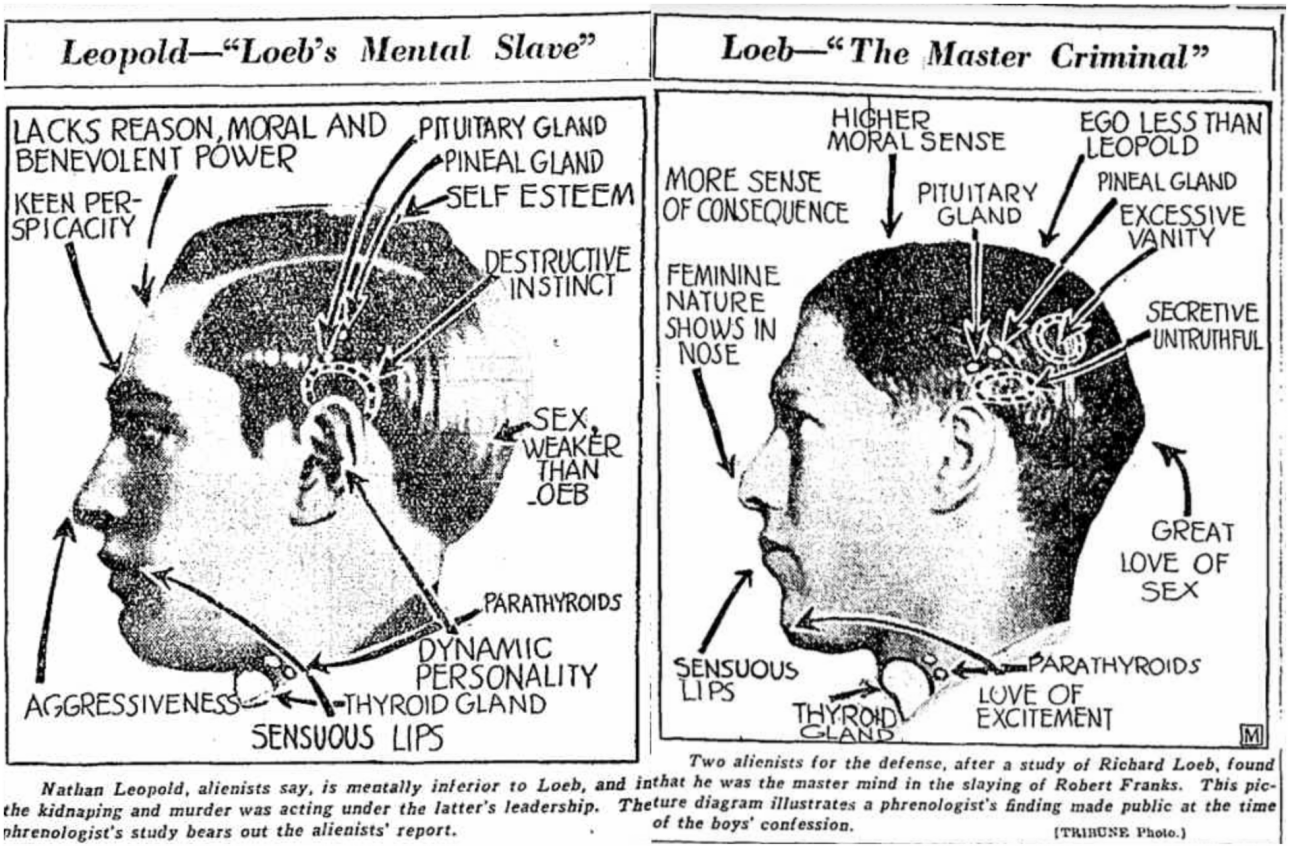


FIGURE 5 | Psychiatrists' ("alienists") phrenological assessments of Mr. Nathan Leopold and Mr. Richard Loeb as they appeared in The Chicago Tribune in the summer of 1924.

neurosciences and our moral and legal reasoning are largely misguided. Even if physicalism holds – as has been assumed here – the neurosciences, by themselves, will not unravel all the physical bases relevant to our psyche and behavior. To accomplish that, the neuroscientific evidence would need to be supplemented by a plethora of other physical information; rather paradoxically, the more physically detailed information of people and their various interactions we gather, the less relevant purely neuroscientific evidence will become. It is clear, therefore, that *neuro*law will never pose a threat to our folk psychological ways of doing moral and legal reasoning. But *physical* law still might. It is reasonable to assume, however, that we will never get there.

CONCLUSION

The mind is dependent, in a crucial way, on its biological basis, the nervous system in particular. Information about this basis should, therefore, have a straightforward impact on our moral and legal reasoning, and, ultimately, on practical jurisprudence. However, despite advances of the neurosciences, neuroscientific evidence has not played a significant role in recent legal cases. Why is that?

Fundamentally, it has here been argued, this is due to the discussion conflating a number of separate issues. As we already know that minds are dependent on brains, finding neural correlates of our psyche and behavior should not come as a surprise to us. Yet, the findings are often portrayed as such. This dualistic – fallacious – sentiment is present also in the discussion on the impact of the neurosciences on jurisprudence. Although we can often point to clear neural changes as being associated with the sort of a behavior, *actus reus*, that is under scrutiny in court proceedings, it is wrong to think that we should conclude that these neural changes are causally responsible for the behavior in question. All behavior has a neural basis, not only the sort that we find morally or legally concerning.

We need, therefore, some independent, and ultimately psychologically and socially based, grounds for thinking that a particular neural change or feature is of such a sort that it should be designated as a cause of some behavior. When deeming a biological basis of decisions and actions dysfunctional, we need to employ psychological and social considerations: it is on the basis of our prior, and often very basic and intuitive psychological and social knowledge that we come to suspect that there is something biologically peculiar in some people, and not the other way around. Only in some rare, very clear cases of externally caused brain lesions are we prone to designate some unequivocally

neural changes as causes of *acti rei*, and to exonerate defendants on the basis such evidence.

Why, then, does neuroscientific evidence of various sorts continue to be presented in court proceedings? Precisely because we are convinced that our psyche and behavior are ultimately neurally based. But even if it were taken for granted, as it here has been, that physicalism holds, and that all our mental states are necessarily dependent on their neural basis, it would be wrong to think it is only neural evidence that we need to rely on to give a complete account of our psyche and behavior. To do that – to completely explain in physical terms some particular mental or behavior features – a more encompassing physical account of the person and her history needs to be given.

We are yet to fully comprehend our nature as thoroughly physical beings in a perfectly physical context. Maybe someday the sciences will paint such a complete picture of us and the surrounding world for us, and maybe that will lead us to abandon the very idea of free will, and the notions of moral and legal responsibility that seem to require such an idea. Whether that is what will indeed happen, is not, however, something that we are in a position to predict. But whatever the verdict will be, it is clear that it is not something that will be reached in some legal process in a particular court of law. It is something that will be reached in the gradual process of all the sciences providing us with a unified understanding of us as conscious, intentional and moral beings.

AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We would like to thank the participants of the Legal Theory Reading Group at the University of Helsinki, the audience of the Fellows Seminar at the Helsinki Collegium for Advanced Studies, Dr. Linus Andersson, four reviewers of *Frontiers in Psychology*, and the Associate Editor Marco Tullio Liuzza for critical comments and discussions. Mr. Erik Rebain provided valuable help on the details pertaining to the case of *The People of the State of Illinois v Nathan F. Leopold Jr. and Richard Loeb*. Part of the work of TKP was funded by the Ella and Georg Ehrnrooth Foundation.

REFERENCES

Alexander, L. (2011). "Culpability," in *The Oxford Handbook of Criminal Law*, eds J. Deigh and D. Dolinko (Oxford: Oxford University Press).

Alexander, L., and Ferzan, K. K. (2009). *Crime and Culpability: A Theory of Criminal Law*. Cambridge, MA: Cambridge University Press.

Alexander, L., and Ferzan, K. K. (2018). *Reflections on Crime and Culpability: Problems and Puzzles*. Cambridge, MA: Cambridge University Press.

Allen, C., and Bekoff, M. (1995). "Function, natural design, and animal behavior: philosophical and ethological considerations," in *Perspectives in Ethology 11: Behavioral Design*, ed. N. S. Thompson (New York, NY: Plenum Press).

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., and Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nat. Neurosci.* 2, 1032–1037. doi: 10.1038/14833

Ashworth, A. (1993). "Taking the consequences," in *Action and Value in Criminal Law*, eds S. Shute, J. Gardner, and J. Horder (Oxford: Oxford University Press).

Ayer, A. J. (1954). *Freedom and Necessity. In His Philosophical Essays*. New York, NY: St Martin's Press.

Bechara, A., Damasio, H., and Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cereb. Cortex* 10, 295–307. doi: 10.1093/cercor/10.3.295

Bechtel, W., and Mundale, J. (1999). Multiple realizability revisited: linking cognitive and neural states. *Philos. Sci.* 66, 175–207. doi: 10.1086/392683

Berofsky, B. (2002). "Ifs, cans, and free will: the issues," in *The Oxford Handbook of Free Will*, ed. R. Kane (Oxford: Oxford University Press).

Block, N. J. (1995). "The mind as the software of the brain," in *Thinking: An Invitation to Cognitive Science*, eds E. E. Smith and D. N. Osherson (Cambridge, MA: The MIT Press).

Boorse, C. (1975). On the distinction between disease and illness. *Philos. Public Affairs* 5, 49–68.

Boorse, C. (1976). What a theory of mental health should be. *J. Theory Soc. Behav.* 6, 61–84. doi: 10.1111/j.1468-5914.1976.tb00359.x

Boorse, C. (1977). Health as a theoretical concept. *Philos. Sci.* 44, 542–573. doi: 10.1086/288768

Boorse, C. (1997). "A rebuttal on health," in *What is Disease?*, eds J. M. Humber and R. F. Almeder (Totowa, NJ: Humana Press).

Braham, M., and van Hees, M. (2012). An anatomy of moral responsibility. *Mind* 121, 601–634.

Brower, M. C., and Price, B. H. (2001). Neuropsychiatry of frontal lobe dysfunction in violent and criminal behaviour: a critical review. *J. Neurol. Neurosurg. Psychiatry* 71, 720–726. doi: 10.1136/jnnp.71.6.720

Burns, J. M., and Swerdlow, R. H. (2003). Right orbitofrontal tumor with pedophilia symptom and constructional apraxia sign. *Arch. Neurol.* 60, 437–440.

Burton, V. S. jr., Cullen, F. T., Evans, D., Alarid, L. F., and Dunaway, R. G. (1998). Gender, self-control, and crime. *J. Res. Crime Delinquency* 35, 123–147.

Cantor, J. M., Kabani, N., Christensen, B. K., Zipursky, R. B., Barbaree, H. E., Dickey, R., et al. (2007). Cerebral white matter deficiencies in pedophilic men. *J. Psychiatric Res.* 42, 167–183. doi: 10.1016/j.jpsychires.2007.10.013

Carbone, J. (2011). "Neuroscience and ideology: why neuroscience can never supply a complete answer for adolescent immaturity," in *Neuroscience and Law: Current Legal Issues*, Vol. 13, ed. M. Freeman (Oxford: Oxford University Press).

Carlisi, C. O., Moffitt, T. E., Knodt, A. R., Harrington, H., Ireland, D., Melzer, T. R., et al. (2020). Associations between life-course-persistent antisocial behaviour and brain structure in a population-representative longitudinal birth cohort. *Lancet Psychiatry* 7, 245–253. doi: 10.1016/s2215-0366(20)30002-x

Carrabine, E., Iganski, P., Lee, M., Plummer, K., and South, N. (2004). *Criminology: A Sociological Introduction*. London: Routledge.

Caruso, G. D. (2012). *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Lanham, MD: Lexington Books.

Caruso, G. D. (2016). Free will skepticism and criminal behavior: a public health-quarantine model. *Southwest Philos. Rev.* 32, 25–48. doi: 10.5840/swphilreview20163214

Caruso, G. D. (2017). *Public Health and Safety: The Social Determinants of Health and Criminal Behavior*. London: ResearchLinks Books.

Caruso, G. D. (2019). A defense of the Luck Pincer: why luck (still) undermines moral responsibility. *J. Infor. Ethics* 28, 51–72.

Chester, D. S., Lynam, D. R., Milich, R., and DeWall, C. N. (2017). Physical aggressiveness and gray matter deficits in ventromedial prefrontal cortex. *Cortex* 97, 17–22. doi: 10.1016/j.cortex.2017.09.024

Chisholm, R. (1964). *Human Freedom and the Self. The Lindley Lecture*. Lawrence, KS: University of Kansas.

Clarke, R. (1993). Toward a credible agent-causal account of free will. *Noûs* 27, 191–203.

Clarke, R. (1996). Agent causation and event causation in the production of free action. *Philos. Top.* 24, 19–48. doi: 10.5840/philtopics19962427

Clarke, R. (2000). Modest libertarianism. *Philos. Perspect.* 14, 21–45. doi: 10.1111/0029-4624.34.s14.2

Clarke, R. (2019). Free will, agent causation, and disappearing agents. *Noûs* 53, 76–96. doi: 10.1111/nous.12206

Damasio, A. R. (1994). *Descartes' Error*. New York, NY: Avon Books.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M., and Damasio, A. R. (1994). The return of Phineas Gage: clues about the brain from the skull of a famous patient. *Science* 264, 1102–1105. doi: 10.1126/science.8178168

Darley, J. M., and Shultz, T. R. (1990). Moral rules: their content and acquisition. *Annu. Rev. Psychol.* 41, 525–556. doi: 10.1146/annurev.ps.41.020190.002521

Darrow, C. (1922). *Crime: Its Cause and Treatment*. New York, NY: Thomas Y. Crowell.

Darrow, C. (1924). "Attorney Clarence Darrow's plea for mercy in the franks case," in *Attorney Clarence Darrow's Plea for Mercy and Prosecutor Robert E. Crowe's Demand for the Death Penalty in the Loeb-Leopold Case*, (Chicago, IL: Wilson Publishing Company).

Darrow, C. (1932). *The Story of My Life*. New York, NY: Grosset and Dunlap.

Davidson, D. (1963). Actions, reasons, and causes. *J. Philos.* 60, 685–700.

Dimock, S. (2012). Intoxication and the act/control/agency requirement. *Crim. Law Philos.* 6, 341–362. doi: 10.1007/s11572-012-9173-x

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., and May, A. (2004). Neuroplasticity: changes in grey matter induced by training. *Nature* 427, 311–312. doi: 10.1038/427311a

Driver, J. (2008a). "Attributions of causation and moral responsibility," in *Moral Psychology The Cognitive Science of Morality*, Vol II, ed. W. Sinnott-Armstrong (Cambridge, MA: The MIT Press).

Driver, J. (2008b). "Kinds of norms and legal causation: reply to Knobe and Fraser and Deigh," in *Moral Psychology The Cognitive Science of Morality*, Vol II, ed. W. Sinnott-Armstrong (Cambridge, MA: The MIT Press).

Driver, J. (2012). *Consequentialism*. Abingdon: Routledge.

Ekstrom, L. (2003). Free will, chance, and mystery. *Philos. Stud.* 113, 153–180.

Elbert, T., Pantev, C., Wienbruch, C., Rockstroh, B., and Taub, E. (1995). Increased cortical representation of the fingers of the left hand in string players. *Science* 270, 305–307. doi: 10.1126/science.270.5234.305

Ellenbogen, J. M., Hurford, M. O., Liebeskind, D. S., Neimark, G. B., and Weiss, D. (2005). Ventromedial frontal lobe trauma. *Neurology* 64:757. doi: 10.1212/wnl.64.4.757

Elzein, N. (2019). "Free will and empirical arguments for epiphenomenalism," in *Causation, Agency, and Supervenience*, Virtues and Economics 5, eds P. Róna and L. Zsolnai (Berlin: Springer).

Elzein, N., and Pernu, T. K. (2017). Supervenient freedom and the free will deadlock. *Disputatio* 9, 219–243. doi: 10.1515/disp-2017-0005

Enoch, D. (2014). "Tort liability and taking responsibility," in *Philosophical Foundations of the Law of Torts*, ed. J. Oberdeek (Oxford: Oxford University Press).

Farahany, N. A. (2016). Neuroscience and behavioral genetics in US criminal law: an empirical analysis. *J. Law and Biosci.* 2, 485–509.

Feinberg, J. (1962). Problematic responsibility in law and morals. *Philos. Rev.* 71, 340–351.

Feinberg, J. (1995). Equal punishment for failed attempts: some bad but instructive arguments against it. *Arizona Law Rev.* 37, 117–133.

Feinberg, J. (2003). *Problems at the Roots of Law: Essays in Legal and Political Theory*. New York, NY: Oxford University Press.

Fischer, J. M., and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge, MA: Cambridge University Press.

Fletcher, G. P. (1998). *Basic Concepts of Criminal Law*. New York, NY: Oxford University Press.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *J. Philos.* 68, 5–20.

Franklin, C. E. (2011a). Farewell to the luck (and Mind) argument. *Philos. Stud.* 156, 199–230. doi: 10.1007/s11098-010-9583-3

Franklin, C. E. (2011b). The problem of enhanced control. *Austr. J. Philos.* 89, 687–706. doi: 10.1080/00048402.2010.524234

Franklin, C. E. (2014). Event-causal libertarianism, functional reduction, and the disappearing agent argument. *Philos. Stud.* 170, 413–432. doi: 10.1007/s11098-013-0237-0

Franklin, C. E. (2018). *A Minimal Libertarianism Free Will and the Promise of Reduction*. New York, NY: Oxford University Press.

Galvani, L. (1791). De viribus electricitatis in motu musculari commentaries. *De Bononiensi Scientiarum et Artium Instituto atque Academia commentarii* 7, 363–418.

Galvani, L. (1794). *Dell'uso e Dell'attività dell'arco Conduttore nelle Contrazioni dei Muscoli*. Bologna: San Tommaso d'Aquino.

Gardner, J. (2001). Legal positivism: 5½ myths. *Am. J. Jurisprudence* 46, 199–227. doi: 10.1093/ajj/46.1.199

Garson, J. (2016). *A Critical Overview of Biological Functions*. Cham, CH: Springer.

Gaser, C., and Schlaug, G. (2003). Brain structures differ between musicians and non-musicians. *J. Neurosci.* 23, 9240–9245. doi: 10.1523/jneurosci.23-27-09240.2003

Ginet, C. (1990). *On Action*. New York, NY: Cambridge University Press.

Ginther, M. R., Sehn, F. X., Bonnie, R. J., Hoffman, M. B., Jones, O. D., Marois, R., et al. (2018). Decoding guilty minds. *Vanderbilt Law Rev.* 71, 241–328.

Glannon, W. (1997). Sensitivity and responsibility for consequences. *Philos. Stud.* 87, 223–233.

Glannon, W. (2002). *The Mental Basis of Responsibility*. Aldershot: Ashgate.

Glannon, W. (2011). "What neuroscience can (and cannot) tell us about criminal responsibility," in *Neuroscience and Law: Current Legal Issues*, ed. M. Freeman (Oxford: Oxford University Press), 13. doi: 10.1093/acprof:oso/9780199599844.003.0002

Goetz, S. T. (2008). *Freedom, Teleology, and Evil*. London: Continuum.

Grafman, J., Schwab, K., Warden, D., Pridgen, A., Brown, H. R., and Salazar, A. M. (1996). Frontal lobe injuries, violence, and aggression: a report of the Vietnam Head Injury Study. *Neurology* 46, 1231–1238.

Greely, H. T. (2009). Law and the revolution in neuroscience: an early look at the field. *Akron Law Rev.* 42, 687–715.

Greely, H. T., and Farahany, N. A. (2019). Neuroscience and the criminal justice system. *Annu. Rev. Criminol.* 2, 451–471.

Greene, J., and Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philos. Trans. R. Soc. Lond. B* 359, 1775–1785.

Griffith, M. E. (2005). Does free will remain a mystery? A response to Van Inwagen. *Philos. Stud.* 24, 261–269. doi: 10.1007/s11098-005-7778-9

Griffith, M. E. (2010). Why agent-caused actions are not lucky. *Am. Philos. Q.* 47, 43–56.

Harlow, J. M. (1848). Passage of an iron bar through the head. *Boston Med. Surg. J.* 13, 389–393. doi: 10.1056/nejm184812130392001

Harlow, J. M. (1868). Recovery from the passage of an iron bar through the head. *Publ. Massachusetts Med. Soc.* 2, 327–347.

Hart, H. L. A., and Honoré, A. (1959). *Causation in the Law*. Oxford: Oxford University Press.

Hatsopoulos, N., Joshi, J., and O'Leary, J. G. (2004). Decoding continuous and discrete motor behaviors using motor and premotor cortical ensembles. *J. Neurophysiol.* 92, 1165–1174. doi: 10.1152/jn.01245.2003

Higdon, H. (1975). *The Crime of the Century: The Leopold and Loeb Case*. New York, NY: G. P. Putnam's Sons.

Hirstein, W., Sifferd, K. L., and Fagan, T. K. (2018). *Responsible Brains: Neuroscience, Law, and Human Culpability*. Cambridge, MA: MIT Press.

Hobbes, T. (1651/1994). *Leviathan: Revised Student Edition. (Cambridge Texts in the History of Political Philosophy)*. Cambridge, MA: Cambridge University Press.

Hume, D. (1738). *A Treatise of Human Nature: An Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. London: John Noon.

Husak, D. N. (1987). *Philosophy of Criminal Law*. Totowa, NJ: Rowman & Littlefield.

Husak, D. N. (1998). "Does criminal liability require an act?," in *Philosophy and the Criminal Law: Principle and Critique,* R. Duff. Cambridge, MA: Cambridge University Press.

Husak, D. N. (2007). Rethinking the act requirement. *Cardozo Law Rev.* 28, 2437–2460.

Husak, D. N. (2011). "The alleged act requirement in criminal law," in *The Oxford Handbook of Philosophy of Criminal Law*, eds J. Deigh and D. Dolinko (Oxford: Oxford University Press).

Jackson, J. B. S. (1870). *A Descriptive Catalogue of the Warren Anatomical Museum*. Boston: A. Williams and Company.

Jacobs, F. G. (1971). *Criminal Responsibility*. London: Weidenfeld and Nicolson.

Jakubiec, M., and Janik, B. (2017). Folk psychology and law: the case of eliminativism. *Semin. Sci.* 16, 155–167.

Jones, O. D., Marois, R., Farah, M. J., and Greely, H. T. (2013). Law and neuroscience. *J. Neurosci.* 33, 17624–17630.

Kane, R. (1996). *The Significance of Free Will*. New York, NY: Oxford University Press.

Kane, R. (1999). Responsibility, luck and chance: reflections on free will and indeterminism. *J. Philos.* 96, 217–240. doi: 10.5840/jphil199996537

Kane, R. (2011). "Rethinking free will: new perspectives on an ancient problem," in *The Oxford Handbook of Free Will*, ed. R. Kane (New York, NY: Oxford University Press).

Khoury, A. C. (2018). The objects of moral responsibility. *Philos. Stud.* 175, 1357–1381.

Kim, J. (1989a). Mechanism, purpose, and explanatory exclusion. *Philos. Perspect.* 3, 77–108.

Kim, J. (1989b). The myth of nonreductive materialism. *Proc. Addresses Am. Philos. Assoc.* 63, 31–47.

Kim, J. (1990). "Explanatory exclusion and the problem of mental causation," in *Information, Semantics and Epistemology*, ed. E. Villanueva (Cambridge, MA: Basil Blackwell).

Kim, J. (1993). "The non-reductivist's troubles with mental causation," in *Mental Causation*, eds J. Heil and A. Mele (Oxford: Clarendon Press).

Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Kim, J. (1999). Making sense of emergence. *Philos. Stud.* 95, 3–36.

Kim, J. (2005). *Physicalism, Or Something Near Enough*. Princeton, NJ: Princeton University Press.

Kim, J. (2007). "Causation and mental causation," in *Contemporary Debates in Philosophy of Mind*, eds B. McLaughlin and J. Cohen (Oxford: Basil Blackwell).

Kingma, E. (2007). What is it to be healthy? *Analysis* 67, 128–133.

Kishiyama, M. M., Boyce, W. T., Jimenez, A. M., Perry, L. M., and Knight, R. T. (2009). Socioeconomic disparities affect prefrontal function in children. *J. Cogn. Neurosci.* 21, 1106–1115. doi: 10.1162/jocn.2009.21101

Koenig-Robert, R., and Pearson, J. (2019). Decoding the contents and strength of imagery before volitional engagement. *Sci. Rep.* 9:3504.

Lagnado, D. A., and Gerstenberg, T. (2017). "Causation in legal and moral reasoning," in *Oxford Handbook of Causal Reasoning*, ed. M. Waldmann (Oxford: Oxford University Press).

Lavergne, G. (1997). *Sniper in the Tower: The Charles Whitman Murders*. Denton, TX: University of North Texas Press.

Lehmann, J., and Gangemi, A. (2007). An ontology of physical causation as a basis for assessing causation in fact and attributing legal responsibility. *Artif. Intell. Law* 15, 301–321. doi: 10.1007/s10506-007-9035-3

Lelling, A. E. (1993). Eliminative materialism, neuroscience and the criminal law. *Univ. Pennsylvania Law Rev.* 141, 1471–1564.

Lemos, J. (2018). *A Pragmatic Approach to Free Will*. New York, NY: Routledge.

Lemos, J. (2020). Kane, Pereboom, and event-causal libertarianism. *Philosophia* 48, 607–623. doi: 10.1007/s11406-019-00098-0

Leopold, N. F. (1963). *Checklist of Birds of Puerto Rico and the Virgin Islands*. San Juan: University of Puerto Rico.

Levy, N. (2008). Bad luck once again. *Philos. Phenomenol. Res.* 77, 749–754. doi: 10.1111/j.1933-1592.2008.00219.x

Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. New York, NY: Oxford University Press.

Levy, N. (2015). Dissolving the puzzle of resultant moral luck. *Rev. Philos. Psychol.* 7, 127–139. doi: 10.1007/s13164-015-0249-0

Lewis, D. (1981). Are we free to break the laws? *Theoria* 47, 113–121. doi: 10.1111/j.1755-2567.1981.tb00473.x

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behav. Brain Sci.* 85, 529–566.

Libet, B. (1994). A testable field theory of mind-brain interaction. *J. Conscious. Stud.* 1, 119–126.

Libet, B. (2002). "Do we have free will?," in *The Oxford Companion to Free Will*, ed. R. Kane (Oxford: Oxford University Press).

Libet, B. (2003). Can conscious experience affect brain activity? *J. Conscious. Stud.* 10, 24–28.

Libet, B. (2004). *Mind Time: The Temporal Factor in Consciousness*. Cambridge, MA: Harvard University Press.

Libet, B. (2006). Reflections on the interaction of the mind and the brain. *Prog. Neurobiol.* 78, 322–326. doi: 10.1016/j.pneurobio.2006.02.003

List, C. (2014). Free will, determinism, and the possibility of doing otherwise. *Noûs* 48, 156–178. doi: 10.1111/nous.12019

List, C. (2019). *Why Free Will Is Real*. Cambridge, MA: Harvard University Press.

List, C., and Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *J. Philos.* 106, 475–502. doi: 10.5840/jphil2009106936

López-Muñoza, F., Boyab, J., and Alamoa, C. (2006). Neuron theory, the cornerstone of neuroscience, on the centenary of the Nobel Prize award to Santiago Ramón y Cajal. *Brain Res. Bull.* 70, 391–405. doi: 10.1016/j.brainresbull.2006.07.010

Macmillan, M. (2000). Restoring Phineas Gate: a 150th retrospective. *J. History Neurosci.* 9, 42–62.

Macmillan, M. (2008). Phineas Gage – unravelling the myth. *Psychol.* 21, 828–831.

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., et al. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proc. Natl. Acad. Sci. U.S.A.* 97, 4398–4403. doi: 10.1073/pnas.070039597

Malle, B. F., Guglielmo, S., and Monroe, A. E. (2014). A theory of blame. *Psychol. Inq.* 25, 147–186.

Mayr, E. (1988). "The multiple meanings of teleological," in *Towards a New Philosophy of Biology*, ed. E. Mayr (Cambridge, MA: Harvard University Press).

McCann, H. J. (1998). *The Works of Agency: On Human Action, Will and Freedom*. Ithaca, NY: Cornell University Press.

Mele, A. R. (1995). *Autonomous agents: From Self-Control to Autonomy*. Oxford: Oxford University Press.

Mele, A. R. (1996). Soft libertarianism and Frankfurt-style scenarios. *Philos. Top.* 24, 123–141. doi: 10.5840/philtopics199624220

Mele, A. R. (2006). *Free Will and Luck*. New York, NY: Oxford University Press.

Menzies, P. (2008). "The exclusion problem, the determination relation, and contrastive causation," in *Being Reduced*, eds J. Hohwy and J. Kallestrup (Oxford: Oxford University Press).

Menzies, P. (2013). "Mental causation in the physical world," in *Mental Causation and Ontology*, eds S. Gibb, E. J. Lowe, and R. Ingthorsson (Oxford: Oxford University Press).

Menzies, P. (2015). The causal closure argument is no threat to non-reductive physicalism. *Hum. .Mente J. Philos. Stud.* 29, 21–46.

Miller, G. A. (2010). Mistreating psychology in the decades of the brain. *Perspect. Psychol. Sci.* 5, 716–743. doi: 10.1177/1745691610388774

Mobbs, D., Lau, H. C., Jones, O. D., and Frith, C. D. (2007). Law, responsibility, and the brain. *PLoS Biol.* 5:e103. doi: 10.1371/journal.pbio.0050103

Moore, G. E. (1903). *Principia Ethica*. Cambridge, MA: Cambridge University Press.

Moore, M. S. (1984). *Law and Psychiatry*. New York, NY: Cambridge University Press.

Moore, M. S. (2009). *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*. Oxford: Oxford University Press.

Moore, M. S. (2020). *Mechanical Choices: The Responsibility of the Human Machine*. Oxford: Oxford University Press.

Morse, S. J. (2003). Inevitable mens rea. *Harvard J. Law Public Policy* 27, 51–64.

Morse, S. J. (2004a). New neuroscience, old problems: legal implications of brain science. *Cerebrum* 6, 81–90.

Morse, S. J. (2004b). The moral metaphysics of causation and results. *Calif. Law Rev.* 88, 879–894.

Morse, S. J. (2006). Brain overclaim syndrome and criminal responsibility: a diagnostic note. *Ohio State J. Crim. Law* 3, 397–412.

Morse, S. J. (2007). Criminal responsibility and the disappearing person. *Cardozo Law Rev.* 28, 2545–2575.

Morse, S. J. (2008). Determinism and the death of folk psychology: two challenges to responsibility from neuroscience. *Minn. J. Law, Sci. Technol.* 9, 1–36.

Morse, S. J. (2011a). "Lost in translation? An essay on law and neuroscience," in *Neuroscience and Law: Current Legal Issues*, ed. M. Freeman (Oxford: Oxford University Press), 13.

Morse, S. J. (2011b). "Neuroscience and the future of personhood and responsibility," in *Constitution 3.0: Freedom and Technological Change*, eds J. Rosen and B. Wittes (Washington, DC: Brookings Institution Press).

Morse, S. J. (2013). Brain overclaim redux. *Law Inequal.* 31, 509–534.

Morse, S. J. (2015). "Neuroscience, free will, and criminal responsibility," in *Free Will and the Brain: Neuroscientific, Philosophical, and Legal Perspectives*, ed. W. Glannon (Cambridge, MA: Cambridge University Press).

Musallam, S., Corneil, B. D., Greger, B., Scherberger, H., and Andersen, R. A. (2004). Cognitive control signals for neural prosthetics. *Science* 305, 258–262. doi: 10.1126/science.1097938

Nozick, R. (1981). *Philosophical Explanations*. Cambridge, MA: Belknap Press.

Nozick, R. (1988). "Knowledge and scepticism," in *Perceptual Knowledge*, ed. J. Dancy (Oxford: Oxford University Press).

O'Connor, T. (1995). "Agent causation," in *Free Will*, ed. Watson (Oxford: Oxford University Press).

O'Connor, T. (2002). *Persons and Causes: The Metaphysics of Free Will*. Oxford: Oxford University Press.

Papineau, D. (1993). *Philosophical Naturalism*. Oxford: Basil Blackwell.

Papineau, D. (2001). "The rise of physicalism," in *Physicalism and Its Discontents*, eds G. Gillett and B. Loewer (Cambridge, MA: Cambridge University Press).

Papineau, D. (2002). *Thinking about Consciousness*. Oxford: Oxford University Press.

Papineau, D. (2009). "The causal closure of the physical and naturalism," in *The Oxford Handbook of Philosophy of Mind*, eds B. McLaughlin, A. Beckermann, and S. Walter (Oxford: Oxford University Press).

Pereboom, D. (2001). *Living Without Free Will*. Cambridge, MA: Cambridge University Press.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.

Pernu, T. K. (2011). Minding matter: how not to argue for the causal efficacy of the mental. *Rev. Neurosci.* 22, 483–507.

Pernu, T. K. (2013a). Does the interventionist notion of causation deliver us from the fear of epiphenomenalism? *Int. Stud. Philos. Sci.* 27, 157–172. doi: 10.1080/02698595.2013.813254

Pernu, T. K. (2013b). The principle of causal exclusion does not make sense. *Philos. Forum* 44, 89–95. doi: 10.1111/phil.12003

Pernu, T. K. (2014a). Causal exclusion and multiple realizations. *Topoi* 33, 525–530. doi: 10.1007/s11245-013-9159-x

Pernu, T. K. (2014b). Interventions on causal exclusion. *Philos. Explorat.* 17, 255–263. doi: 10.1080/13869795.2013.805800

Pernu, T. K. (2016). Causal exclusion and downward counterfactuals. *Erkenntnis* 81, 1031–1049. doi: 10.1007/s10670-015-9781-7

Pernu, T. K. (2017). The five marks of the mental. *Front. Psychol.* 8:1084. doi: 10.3389/fpsyg.2017.01084

Pernu, T. K. (2018). Mental causation via neuroprosthetics? A critical analysis. *Synthese* 195, 5159–5174. doi: 10.1007/s11229-018-1713-z

Polger, T. W., and Shapiro, L. A. (2016). *The Multiple Realization Book*. Oxford: Oxford University Press.

Ponseti, J., Siebner Hartwig, R., Klöppel, S., Wolff, S., Granert, O., Jansen, O., et al. (2007). Homosexual women have less grey matter in perirhinal cortex. *PLoS One* 2:e762. doi: 10.1371/journal.pone.0000762

Raatikainen, P. (2010). Causation, exclusion, and the special sciences. *Erkenntnis* 73, 349–363. doi: 10.1007/s10670-010-9236-0

Raine, A., Buchsbaum, M. S., Stanley, J., Lottenberg, S., Abel, L., and Stoddard, J. (1994). Selective reductions in prefrontal glucose metabolism in murderers. *Biol. Psychiatry* 36, 365–373. doi: 10.1016/0006-3223(94)91211-4

Raine, A., Lencz, T., Bihrle, S., LaCasse, L., and Colletti, P. (2000). Reduced prefrontal gray matter volume and reduced autonomic activity in antisocial personality disorder. *Arch Gen Psychiatry* 57, 119–127.

Ramón y Cajal, S. (1888). Estructura de los centros nerviosos de las aves. *Revista Trimestral de Histología Normal y Patológica* 1, 1–10.

Reid, T. (1788/1969). *Essays on the Active Powers of the Human Mind*. Cambridge, MA: MIT Press.

Rosen, J. (2007). *The Brain on the Stand*. New York, NY: The New York Times, 11.

Rowe, D., Vazsonyi, A., and Flannery, D. (1995). Sex differences in crime: do means and within-sex variation have similar causes? *J. Res. Crime Delinquency* 32, 84–100. doi: 10.1177/0022427895032001004

Sajous-Turner, A., Anderson, N. E., Widdows, M., Nyalakanti, P., Harenski, K., Harenski, C., et al. (2019). Aberrant brain gray matter in murderers. *Brain Imaging Behav.* 5:10.1007/s11682-019-00155-y. doi: 10.1007/s11682-019-00155-y

Sapolsky, R. M. (2004). The frontal cortex and the criminal justice system. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 359, 1787–1796.

Sartorio, C. (2007). Causation and responsibility. *Philosophy Compass* 2, 749–765.

Sartorio, C. (2016). *Causation and Free Will*. Oxford: Oxford University Press.

Savic, I., and Lindström, P. (2008). PET and MRI show differences in cerebral asymmetry and functional connectivity between homo- and heterosexual subjects. *Proc. Natl. Acad. Sci. U.S.A.* 105, 9403–9408. doi: 10.1073/pnas.0801566105

Schleim, S., Spranger, T. M., Erk, S., and Walter, H. (2011). From moral to legal judgment: the influence of normative context in lawyers and other academics. *Soc. Cogn. Affect. Neurosci.* 6, 48–57. doi: 10.1093/scan/nsq010

Schlick, M. (1930). *Fragen der Ethik*. Wien: Verlag von Julius Springer.

Schneider, B., and Koenings, M. (2017). Human lesion studies of ventromedial prefrontal cortex. *Neuropsychologia* 107, 84–93. doi: 10.1016/j.neuropsychologia.2017.09.035

Shapiro, L. A. (2000). Multiple realizations. *J. Philos.* 97, 635–654.

Shaw, E., Pereboom, D., and Caruso, G. D. (eds) (2019). *Free Will Skepticism in Law and Society: Challenging Retributive Justice*. New York, NY: Cambridge University Press.

Shenoy, K. V., Meeker, D., Cao, S., Kureshi, S. A., Pesaran, B., Buneo, C. A., et al. (2003). Neural prosthetic control signals from plan activity. *NeuroReport* 14, 591–596. doi: 10.1097/00001756-200303240-00013

Shultz, T. R., and Schleifer, M. (1983). "Towards a refinement of attribution concepts," in *Attribution Theory and Research: Conceptual, Developmental and Social Dimensions*, eds J. Jaspers, F. D. Fincham, and M. Hewstone (London: Academic).

Sifferd, K. L. (2006). In defense of the use of commonsense psychology in the criminal law. *Law Philos.* 25, 571–612. doi: 10.1007/s10982-005-3802-7

Sifferd, K. L. (2018). "Non-eliminative reductionism: not the theory of mind some responsibility theorists want, but the one they need," in *Neurolaw and Responsibility for Action: Concepts, Crimes, and Courts*, ed. B. Donnelly-Lazarov (Cambridge, MA: Cambridge University Press).

Simester, A. P. (2017). Causation in (criminal) law. *Law Q. Rev.* 133, 416–441.

Sloman, S. A., Fernbach, P. A., and Ewing, S. (2009). "Causal models: the representational infrastructure for moral judgment," in *Moral Judgment and Decision Making (Psychology of Learning and Motivation*, Vol. 50, eds D. Bartels, C. Bauman, L. Skitka, D. L. Medin, and B. H. Ross (San Diego, CA: Academic Press).

Smart, J. J. C. (1961). Free will, praise and blame. *Mind* 70, 291–306.

Sober, E. (1980). Evolution, population thinking and essentialism. *Philos. Sci.* 47, 350–383. doi: 10.1086/288942

Soon, C. S., Brass, M., Heinze, H.-J., and Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nat. Neurosci.* 11, 543–545. doi: 10.1038/nn.2112

Sterr, A., Müller, M. M., Elbert, T., Rockstroh, B., Pantev, C., and Edward, T. (1998). Changed perceptions in Braille readers. *Nature* 391, 134–135. doi: 10.1038/34322

Steward, H. (2012). *A Metaphysics for Freedom*. Oxford: Oxford University Press.

Szigeti, A. (2014). "Collective responsibility and group-control," in *Rethinking the Individualism-Holism Debate*, eds J. Zahle and F. Collin (Cham: Springer).

Tancredi, L. R., and Brodie, J. D. (2007). The brain and behaviour: limitations in the legal use of functional magnetic resonance imaging. *Ame. J. Law Med.* 271, 288–289.

Taylor, R. (1966). *Action and Purpose*. Englewood Cliffs: Prentice-Hall.

Texas Governor's Committee and Consultants (1966). *Report on the Charles J. Whitman Catastrophe*. Austin, TX: Texas State Library and Archives Commission.

Theodore, J. (2007). *Evil Summer: Babe Leopold, Dickie Loeb, and the Kidnap-Murder of Bobby Franks*. Carbondale, IL: Southern Illinois University Press.

van Elst, L. T., Woermann, F. G., Lemieux, L., Thompson, P. J., and Trimble, M. R. (2000). Affective aggression in patients with temporal lobe epilepsy: a quantitative MRI study of the amygdala. *Brain* 123, 234–243. doi: 10.1093/brain/123.2.234

Vihvelin, K. (2004). Free will demystified: a dispositional account. *Philos. Top.* 32, 427–450. doi: 10.5840/philtopics2004321/211

Vihvelin, K. (2011). "How to think about the free will/determinism problem," in *Carving Nature at its Joints*, eds J. K. Campbell and M. O'Rourke (Cambridge, MA: MIT Press).

Vihvelin, K. (2013). *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. New York, NY: Oxford University Press.

Volkow, N. D., and Tancredi, L. (1987). Neural substrates of violent behaviour. A preliminary study with positron emission tomography. *Br. J. Psychiatry* 151, 668–673. doi: 10.1192/bjp.151.5.668

Voss, M., Moore, J., Hauser, M., Gallinat, J., Heinz, A., and Haggard, P. (2010). Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences. *Brain* 133, 3104–3112. doi: 10.1093/brain/awq152

Walker, J. T., and Maddan, S. (2013). *Understanding Statistics for the Social Sciences, Criminal Justice, and Criminology*. Burlington, MA: Jones & Bartlett Learning.

Walker, N. (1968). *Crime and Insanity in England. One: The Historical Perspective*. Edinburgh: Edinburgh University Press.

Waller, B. N. (1990). *Freedom Without Responsibility*. New York, NY: Temple University Press.

Waller, B. N. (2011). *Against Moral Responsibility*. Cambridge, MA: MIT Press.

Watson, G. (1975). Free agency. *J. Philos.* 71, 205–220.

Wegner, D. M. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wegner, D. M. (2004). Précis of The Illusion of Conscious Will. *Behav. Brain Sci.* 27, 649–659.

Weiss, K. J. (2011). Head, examined: clarence Darrows x-ray vision of criminal responsibility. *J. Psychiatry Law* 39, 627–661. doi: 10.1177/009318531103900406

Whittle, A. (2018). Responsibility in context. *Erkenntnis* 83, 163–183.

Willemsen, P. (2019). *Omissions and Their Moral Relevance: Assessing Causal and Moral Responsibility for the Things We Fail to Do*. Leiden, NL: Mentis Verlag.

Wilson, D. (2015). *Genetics, Crime and Justice*. Cheltenham: Edward Elgar Publishing.

Woermann, F. G., van Elst, L. T., Keopp, M. J., Free, S. L., Thompson, P. J., Trimble, M. R., et al. (2000). Reduction of frontal neocortical grey matter associated with affective aggression in patients with temporal lobe epilepsy: an objective voxel by voxel analysis of automatically segmented MRI. *J. Neurol. Neurosurg. Psychiatry* 68, 162–169. doi: 10.1136/jnnp.68.2.162

Wolf, S. (1980). Asymmetrical freedom. *J. Philos.* 77, 151–166.

Wolf, S. (1990). *Freedom Within Reason*. New York, NY: Oxford University Press.

Woodward, J. F. (2008). "Mental causation and neural mechanisms," in *Being Reduced: New Essays on Reductive Explanation and Special Science Causation*, eds J. Hohwy and J. Kallestrup (Oxford: Oxford University Press).

Yang, Y. L., Raine, A., Lencz, T., Bihrle, S., Lacasse, L., and Colletti, P. (2005). Prefrontal white matter in pathological liars. *Br. J. Psychiatry* 187, 320–325. doi: 10.1192/bjp.187.4.320

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

REVIEW
published: 27 June 2019
doi: 10.3389/fpsyg.2019.01406

Criminal Responsibility and Neuroscience: No Revolution Yet

Ariane Bigenwald[1,2]* *and Valerian Chambon*[2]*

[1] *Département de Philosophie, Université Paris I Panthéon Sorbonne, Paris, France,* [2] *Institut Jean Nicod (ENS – EHESS – CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL University, Paris, France*

Since the 1990's, neurolaw is on the rise. At the heart of heated debates lies the recurrent theme of a neuro-revolution of criminal responsibility. However, caution should be observed: the alleged foundations of criminal responsibility (amongst which free will) are often inaccurate and the relative imperviousness of its real foundations to scientific facts often underestimated. Neuroscientific findings may impact on social institutions, but only insofar as they also engage in a political justification of the changes being called for, convince populations, and take into consideration the ensuing consequences. Moreover, the many limits of neuroscientific tools call for increased vigilance when, if ever, using neuroscientific evidence in a courtroom. In this article, we aim at setting the basis for future sound debates on the contribution of neuroscience to criminal law, and in particular to the assessment of criminal responsibility. As such, we provide analytical tools to grasp the political and normative nature of criminal responsibility and review the current or projected use of neuroscience in the law, all the while bearing in mind the highly publicized question: can neuroscience revolutionize criminal responsibility? Answering this question implicitly requires answering a second question: *should* neuroscience revolutionize the institution of criminal responsibility? Answering both, in turn, requires drawing the line between science and normativity, revolution and dialogue, fantasies and legitimate hopes.

Keywords: criminal responsibility, liability, free will, sense of agency, neuroscience, neurolaw, cognitive bias, moral agent

OPEN ACCESS

Edited by:
Eric García-López, Instituto Nacional de Ciencias Penales, Mexico

Reviewed by:
Andrea Lavazza, Centro Universitario Internazionale, Italy
Jose Angel Marinaro, National University of La Matanza, Argentina

***Correspondence:**
Ariane Bigenwald
ariane.bigenwald@gmail.com
Valerian Chambon
valerian.chambon@gmail.com

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *15 February 2019*
Accepted: *31 May 2019*
Published: *27 June 2019*

Citation:
Bigenwald A and Chambon V (2019) Criminal Responsibility and Neuroscience: No Revolution Yet. Front. Psychol. 10:1406. doi: 10.3389/fpsyg.2019.01406

INTRODUCTION

"A truly scientific, mechanistic view of the nervous system make[s] nonsense of the very idea of responsibility" states Dawkins, a biologist for whom neuroscience would overthrow the retributivist foundations of criminal law (Dawkins, 2006). Others abound in his direction: supporting that "free will is an illusion," Greene and Cohen (2004) wish to replace retribution with deterrence, prevention and medical treatment. In a similar vein, Sapolsky upholds "a world of criminal justice in which there is no blame, only prior causes" (Sapolsky, 2004).

Neuroscience and indeed all disciplines studying brain structure and function have had a growing influence on political discourse, particularly in the legal sphere. Since the 1990's, "neurolaw" has emerged as a new cross-disciplinary field of study. At the heart of heated debates lies the recurrent question of criminal responsibility, and an enthusiasm, just as recurrent, for an alleged overthrow of this notion by the fast-growing discipline of brain sciences. The theme of a neuro-revolution is indeed popular in the media and scientific and philosophical literature.

However, the link between science and law – between the explanatory and the normative – is far from self-evident, and the ties between neuroscience and criminal responsibility are still far from convincing. The alleged, and supposedly challenged, foundations of criminal responsibility (not least of which is the notion of free-will) are not only wrong. The real foundations of responsibility, embedded as they are in our daily experiences and ideological framework, are relatively impervious to scientific facts. They are susceptible to the latter, but only insofar as these may constitute an argument in favor of a political or ideological alternative. Moreover, the many limits (e.g., technical, interpretative, etc.) of neuroscientific tools and measurements call for increased vigilance when, if ever, using neuroscientific evidence in a courtroom.

Can neuroscience revolutionize criminal responsibility? Answering this question implicitly requires answering a second question: *should* neuroscience revolutionize the institution of criminal responsibility? Answering both, in turn, requires drawing the line between science and normativity, revolution and dialogue, fantasies and legitimate hopes. We aim here to introduce those nuances. In order to do so, we will first define criminal responsibility and elaborate on the principles and normativity behind this model. We will then address the limits to using neuroscience in the courts. Finally, we will evaluate the concrete and more modest contributions of neuroscience to the judicial process.

BOX 1 | Criminal responsibility.
Criminal responsibility is based on the *actus reus* and the *mens rea.* To be criminally liable, one must thus (1) consciously will to *x*; (2) know that *x* is wrong; and (3) do *x.* The presence of neurological prior causes to that action, or the predictability of an action due to identified priors, is a matter that relates to free will (how does one forms intentions? where do they come from? etc.). Responsibility, on the other hand, only cares for the feeling of consistency in the causal chain between intention and effect (*intention-action-effect* chain). What judges evaluate is the accused's capacity to act in accordance with his or her intentions. The accused's narrative on his or her agency is then normatively evaluated: that is, the narrative is confronted to current common beliefs and values. If you would report having intentionally killed your neighbor while knowing that it was wrong at the time you did it, but add that you did so following Satan's orders, you would not be considered liable for your acts because you don't share the Law's normative reality: a secular reality in which Satan does not exist. Criminal responsibility, hence, lies in the individual's subjective experience of agency and on the normative assessment of that experience.

WHAT IS CRIMINAL RESPONSIBILITY?

Brief Definition and Basic Legal Principles

Before getting to the heart of the matter, some preliminary definitions are needed, especially regarding the definition of responsibility. As with any ambiguous term, "responsibility" allows for several meanings[1]: a tree falling on an electrical wire can be said to be responsible for a power failure (causal meaning), the captain of a ship is responsible for safety on board (role), a young man can be particularly irresponsible (character), insurers are responsible for compensating road accident victims (civil liability), a patient can be diagnosed irresponsible by psychiatrists (capacity), I can be responsible of my own misfortunes (authorship, or practical meaning), and so on. Criminal responsibility mixes different meanings (practical and capacity), but applies especially to social and legal norms (normative meaning). More specifically, a person is *prima facie* criminally responsible when he or she commits a crime while validating its constitutive elements: the *actus reus* and the *mens rea* (**Box 1**). The *actus reus* is the material element of a crime, which is to say the act that is being reprimanded, and the *mens rea* is the mental element, which is to say the state of mind of the accused at the moment of committing that act. A murder, for example, requires both the act of killing a person and the specific intent of killing that person. Without this *mens rea*, the act of killing someone does not amount to murder, but manslaughter. *Mens rea* is evaluated either subjectively through intention, carelessness or wilful blindness, or objectively, in comparison with a "reasonable person" facing similar circumstances, through negligence or recklessness. The elements required to prove those states of the mind are knowledge (of the nature of the act, of its consequences and of surrounding circumstances) and will (in the sense of a wilful act, i.e., an act that is part of a conscious plan of action). All of these terms have the same meaning as in ordinary language.

To understand the scope of criminal responsibility, it is also important to grasp its limits, and hence, the classification of legal defenses. In Canadian criminal law, for example[2], defenses are traditionally divided into two categories, and relate to situations affecting the capacity to orient one's actions either "cleverly (intelligently)" or "freely"[3] . The first is composed of factors such as minor status, mental illness, automatism, intoxication, and error, while the latter includes necessity, coercion, provocation, impossibility, and self-defense. Briefly put, we excuse the incompetent: those who cannot understand or could not act. In ordinary language, this corresponds to the distinction between excuse and justification. An excuse is exculpatory because it casts doubt on the presence of *mens rea.* Justification, on the other hand, is a mitigating factor that reduces either the infraction or the sentence, since it intervenes after the *actus reus* and *mens rea* have been proven.

This definition of criminal responsibility outlines a particular vision of the responsible agent. *Mens rea,* as well as the typology

[1]H.L.A. Hart, a renowned philosopher of law, emphasized this ambiguity in a famous excerpt on a captain of a ship responsible (or not responsible) in so many ways of so many different things (see Hart, 1968, p. 211).

[2]Hereafter we will often take the model of Canadian Criminal law – which we are most familiar with –, but its structure is also present in a number of other legal systems. The need for both a material and a psychological element of the crime to define responsibility, as well as the main exonerations, are indeed found in most, if not all, of the western-inspired legal systems. To have a glance at national criminal law systems, see: International Encylopedia for Criminal Law, ed. by Dr. Frank Verbruggen and Dr. Vanessa Franssen, online : http://www.kluwerlawonline.com/toc.php?pubcode=CRIM. We also note that international criminal law is built upon the same basic categories of material and psychological elements of the crime and exonerations such as mental disorder, intoxication, necessity, duress, self-defense, etc. (see: art. 30 and 31 of the Rome Statute).

[3]Such a categorisation is borrowed from Parent (2008).

of excuses, reflects the expectations we have of finding certain capacities in the ones we judge. In this regards, it can be said that criminal law is "capacitarian" (Vincent, 2010). Responsible agents are thus individuals capable of orienting their actions intentionally, consciously, and more or less rationally in a manner suitable to the normative framework in which they act. Besides, they must not be coerced into violating that framework.

All of those criteria are evaluated according to the individual's behavior. The *causes* of behavior are not taken into consideration. In other words, a person is excused on the basis of an automatism, for example, whether caused by a physical or supernatural phenomenon.

Why Free Will Does Not Matter

In this section, we take the liberty to venture into the question of free will, as it continues to haunt neuroscientific discourse on responsibility. However, it is essential to take this elusive question for what it is: a ghost – a dead specter that resurfaces when it is not properly put to rest.

That discourse considers the foundation of responsibility to be free will, taken in a general sense as meaning the possibility of "avoiding wrongdoing" or of "acting otherwise"[4]. The notion of determinism, as put forward by neuroscience, by reducing each of our actions to their neurological and unconscious causes, and therefore treating them as mere events rather than wilful actions, would appear to render the possibility of alternative outcomes illusory. Consequently, we would not be responsible, unless some other notion could be identified to salvage human agency and thus, responsibility itself.

Admittedly, this is a grossly simplistic definition of determinism. The reason for this approximation is the ongoing dissension over the central notion of causality between scientists and philosophers (e.g., Frisch, 2014, for a review). In the interest of being inclusive, we will therefore refer to *determinisms*. One should also notice that debates surrounding free will and determinisms are metaphysical, hence arguing within and opposing different ontologies. As we hope to show, there is no need for solving such metaphysical debates.

Having outlined these precautions, we can now turn to the debate on responsibility, which is distinct from free will and practical in nature. In other words, criminal responsibility is not founded in free will but on practical, subjective and political considerations.[5] As such, it is impervious to any truth about determinisms.

First, determinisms alone, even if true, do not annihilate the feeling that I have of controlling my actions. Indeed, I always have the luxury of contradicting anyone's predictions on my behavior (Searle, 1984). This "subjectivist" objection, defended by Searle and others (Chisholm, 1976; Baertschi, 2009), is not to be taken for an argument against determinisms. It is rather an argument in favor of our current concept of a responsible agent. Not only would this argument promulgate a wrong definition of determinism (Russell, 1912), but it mostly does not seek to address determinism at all. It does not matter whether or not my intentions, my feeling of control, my actions and their results are predetermined, or "caused," by non-conscious antecedents – such as preparatory activity in subcortical and frontal motor areas (e.g., Cunnington et al., 2003). The argument consists of emphasizing that what does matter is the power to *associate* a conscious will to an act or to the results thereof. Therefore, our actions need only conform to our intentions and be perceived as part of a conscious plan of action, i.e., a plan that integrates an explicit and non-ambiguous representation of the action's potential consequences (see Synofzik et al., 2008). The institution of responsibility thus lies in the possibility for the individual to experience *agency*, a subjective feeling of being causally responsible for his or her actions and consequences thereof (Haggard and Chambon, 2012) (see **Box 2**, "Sense of agency"). Furthermore, determinisms, even if they were proven to be true and deeply anchored in agents' beliefs, do not modify the conscience about what is appropriate and what is not, what is deemed socially acceptable or unacceptable. Determinisms do not forbid the promulgation of norms. Knowing that I can always act in conformity with my intention and still tell apart right from wrong, I can more often than not decide to act rightfully, or at least believe that I am. In this respect, criminal responsibility relies mostly on our subjective experience, the impression of being able to choose to act or avoid acting.[6]

Some legal and popular expressions may lead us to think that responsibility is nonetheless grounded in free will. Everyone legitimately assumes, for example, that criminal proceedings aim at evaluating if the accused "could have acted otherwise." H.L.A. Hart, famous legal philosopher, takes the "fair chance of avoiding wrongdoing" to be the foundation of criminal responsibility. However, our previous paragraph has already shown the subjective interpretation plausibly given to this standard: the accused only needs to have had the impression of being able to avoid wrongdoing. The political interpretation of that standard shall now cast away free will once and for all. What matters most in the principle of a "fair chance of avoiding wrongdoing" is the word "fair" rather than "chance," and the word "fair" taken as meaning equitable rather than just. It is not about knowing if there was actually a chance – another possible course of action – but rather if the circumstances as offered by society or a given social environment, were equitable and favorable to the expression of a singular "conscience" through a choice. By analyzing *mens rea,* the judge does not wish to know if the accused could have acted otherwise, but if the circumstances surrounding the crime were preventing awareness, and a sense

[4]Note that the definition of free will is contentious in itself. According to Frankfurt, an agent is "free" if he wants what he wants, such that his lower-order desires correspond to higher-order volitions (e.g., Frankfurt, 1988). For others (Descartes, Berkeley, Kant), free will requires that an agent can genuinely escape the causal necessity of a deterministic world.

[5]The debate on criminal responsibility is independent from the discussion on (moral) naturalism, that is from knowing whether the law, as a system of established rules, derives from our moral *intuitions* or not. We also recall that we only deal with criminal responsibility, without delving into drawing the lines of intersection between law and morality, e.g., assessing the question of whether or not moral responsibility is also independent from free will or if it is intertwined with criminal responsibility.

[6]As the next section (and next paragraph) will show, assessment of this subjective experience is also normative.

BOX 2 | Sense of agency.

Agency refers to an individual's capacity to initiate and perform actions, and thus to bring about change, both in their own state, and in the state of the outside world (Chambon et al., 2014a). Thus, agency is an objective fact, demonstrated by individuals' behaviors and the consequences of those behaviors. But agency has a first-component as well: it involves a subjective experience unique to the agent. The *experience* of agency, also referred to as "sense of agency," is classically defined as a phenomenal experience of "mineness" of one's own action (Synofzik et al., 2008; Eitam and Haggard, 2015). Whether this (minimal) action-related self-awareness relies on a *post hoc* cognitive *reconstruction*, or relies on internal signals being experienced while preparing and executing the action (e.g., Chambon et al., 2014b), is of little relevance for judging criminal responsibility. Criminal responsibility is acknowledged when, together with the material element of a crime (actus reus), criteria for *subjective* agency (whether the agent is sensing, experiencing, or reporting to have some sort of authorship over an action) are met.

Making a choice vs. Having a choice

Shepard and O'Grady criticize the univocal use of "choice" in folk psychology (Shepard and O'Grady, 2017). In a recent empirical work, they show that there are at least two distinct, though related, concepts of choice: one expressed in the phrase 'making a choice' and another expressed in the phrase 'having a choice.' "*One difference between these concepts,"* argue the authors*, "involves the kinds of alternatives each is sensitive to. Making a choice is primarily sensitive to whether or not psychologically open alternatives are present and whether an agent's decision goes through normal psychological processes, but only minimally sensitive to whether or not genuinely open alternatives are present (. . .) In contrast, having a choice is sensitive to whether genuinely open alternatives are present, and whether psychologically open alternatives are present*". Shepard and O'Grady relate this conceptual difference to a judgment on free will (which in turn they relate to responsibility). While they acknowledge that only few studies have investigated this link between choice and free will, with conflicting results, they note that "*findings suggest attributions of free will more closely mirror attributions of making a choice than having a choice*" (see also Shepard and Reuter, 2012; Nahmias and Thompson, 2014; Nahmias et al., 2014). The conceptual distinction between "having a choice" and "making a choice" echoes another distinction between causal responsibility and criminal responsibility, which we mentioned above. Thus, studies showing that the number and the availability of alternatives (*counterfactuals*) influence judgments on causal responsibility (Kulakova et al., 2017) are not directly relevant to determine what influences judgments of moral and criminal responsibility (Shaver, 2012).

of authorship, of the act.[7] Society assesses the fairness of the conditions it gave to the accused that it is judging. Finally, this equity is based on the subjective belief that we consciously guide our actions according to our own motivations and on the collective belief that we are, in fact, endowed with motivations in the first place.

Anecdotally, it may be added that this observation is echoed in folk psychology. Recent studies have revealed the possibility that blaming might depend less on the availability of actual open alternatives than on the availability of psychologically open alternatives, i.e., that blame might be based on the appreciation of the accused's subjective belief of having made a choice (Shepard and O'Grady, 2017). Those studies have demonstrated that there is a conceptual difference between "having a choice" and "making a choice," and that it is possible that the second category is more relevant to the act of judging one's responsibility (see **Box 2**, "Making a choice *vs*. Having a choice").

One shall recall the great attention paid to Libet's famous experiment and Wegner's illusionism (Libet, 1999; Wegner, 2002). Following Libet's results showing that a certain brain activity related to conscious actions systematically preceded the agent's conscious intention, multiple interpretations suggested that conscious will was not the cause of our actions[8], that we had no free will, and that we therefore could not possibly be responsible (Sinnott-Armstrong and Nadel, 2010). Without commenting on the validity of such theses (see Schurger et al., 2012; Frith and Haggard, 2018), it is obvious from our previous analysis that these do not impact criminal responsibility. They might have had an impact if criminal responsibility were based on free will (and in this case, more specifically, on the absence of neurological prior causes). However, as we have already pointed out, this is not the case. As long as the illusion of free will remains intact, even if it is an illusion, we can claim to be responsible. The responsible agent is only required to have an internal plan of action, including a representation of the planned behavior (intention), and to have sufficient insight into the normally possible consequences of that behavior (knowledge) (Synofzik et al., 2008). In this regard, the origins of an intention do not matter. What criminal responsibility requires is an individual's capacity to act in a manner deemed appropriate to the realization of the related intention, given his or her knowledge of social norms defining what is acceptable and unacceptable.

UNDERSTANDING NORMATIVITY: NEUROSCIENCE TELLS BUT DOES NOT COMPEL

Responsibility is immune from determinisms not only by virtue of its independence from free will. In fact, no scientific discovery, as significant as it may be, in and of itself calls for the overthrow or modification of a social institution. In other words, we insist on the difference between positive and normative, also called the 'is-ought gap,' and will explain further the particularities of normativity.

The Morse Challenge

Hume brought the irreducibility of *what is* to *what should be* to light in the XVIII century. The idea goes as follows: nothing that simply is calls directly for what should be, *without postulating that "what should be" (what is good) should be conform to what is.* At the junction of law and neuroscience, S. J. Morse reaffirmed the Humean argument to defeat naively enthusiastic scientific claims in courtrooms. In his famous article "*Brain overclaim syndrome*

[7] In this regard, one notes that criminal law contextualizes the accused's actions "in the ordinary course of events" – what "ordinary" means being left to a political (social, cultural, etc.) appreciation.

[8] In this case, the interpretation comes from Wegner himself. The logic is the following: since it has been observed that a pre-motor potential (or readiness potential) occurs about 600 ms before conscious awareness of intention, which in turn occurs about 200 ms before action onset, the belief that we intentionally cause our actions (in other words that "consciously willing the action" causes the initiation of action) would be an illusion.

and responsibility: a diagnostic note" (Morse, 2006), he recalls the behavioral, as opposed to cerebral, criteria for responsibility and insists on the incapacity of brain imaging to set the threshold of normality vs. abnormality either in ethics or in law. "*Brains are not responsible. Acting people are*" (p. 406)[9]. Hence, explaining the difference in behaviors between a teenager and an adult by the lack of complete myelinization of cortical neurons as in Roper v. Simmons (2005)[10], and inferring as a result the lack of sufficient responsibility to qualify for the death penalty, is simply irrelevant (p. 397).[11] It only takes a difference in behavior between those two types of individuals. Baerstchi complements the "Morse challenge" by showing concrete Humean limits in some experiments (Baertschi, 2009). Some studies have outlined the different brain areas which operate in the course of moral decision-making, when faced with the well-known trolley dilemma (Roskies, 2004). Those areas, while of interest to indicate the part played by emotions in moral decisions, do not inform the manner, consequentialist or deontologist, in which to settle this dilemma.

Responsibility Is a Normative Concept

The requirements for responsibility are normative, which is to say that they are standards that claim to originate in a social choice and to have practical authority. These norms are guided by beliefs and principles.

For example, the legal principle of non-retroactivity: according to this principle, it is fair to be judged only by laws that you had the opportunity to know about before committing an offense. This principle implies that individuals are capable, or believe themselves to be capable, of orienting their actions so as to avoid negative consequences (here, a criminal penalty). One way to take this principle and the belief it implies into account is to establish the state of mind of the accused at the time of the events. Considering, as highlighted above, that *mens rea* also serves the purpose of ensuring the equity of the circumstances in which the accused acted and thus ensure that he or she had a "fair chance of avoiding wrongdoing."

The responsible agent's abilities, such as intentionality and rationality, are also normative. Phineas Gage is a classical example. In this case, a man who suffered great brain lesions after an accident started to adopt negative behaviors. When thinking abstractly, he could make a good decision, but, when facing a concrete situation, he would systematically make a bad one. However, when deeming his behavior as good or bad, we already interpret his actions according to a normative standard of rationality (Baertschi, 2009). Gage was incapable of reasoning about a decision directly related to him or his personal circle, of acting rationally according to his best interest (whatever definition of interest is taken). We then consider that he lacks an essential characteristic of practical rationality, i.e., the ability to apply logical reasoning to a concrete objective deemed beneficial. Once more, this conclusion relies on a common definition of rationality and does not rely on Gage's brain injury.

Another example of normativity at work in responsibility assignments concerns "reality" itself. For some God exists, for others he does not. Depending on whether we are atheists or believers, "God has asked me to do it" is either a madman's whim or a saint's word. The difference between the madman and the saint is not so much a question of belief than it is a question of norms and society. The madman is a saint if we share his reality, and the saint a madman if we don't. An implicit norm is thus at work in any legal judgment, as minimally relating to reality. Our beliefs are involved in what we deem rational. What we recognize to be rational is partly arbitrary, precisely because we recognize it.

In the previous section, we insisted on the experiential requirement: the accused must be able to report a feeling of agency to potentially be responsible. We added in this section another criterion: responsibility also depends on a normative appreciation of that subjective experience, i.e., a normative attribution of agency (of what we commonly call agency).

Changing the Premises of Responsibility Is a Social Decision

To be efficient at an institutional level and in order to inform juridical considerations, neuroscience must accept that scientific facts alone are not enough, and that these must be integrated into a broader normative scheme if they are to have any legal significance. It must convince us beyond and against our daily experiences that our rationality is sufficiently flawed, that our will is powerless, that our choices are all about neurological prior causes, to the point that we should doubt everything we are told by this "rationality," this "will" and those "choices," etc. It must acquire normative authority. After all, why not? Ancient Greeks certainly did not have the same individualistic appreciation of the artist's agency: the writer would simply copy words dictated by muses. Neuroscience would nonetheless be leaving the field of science for the bumpier grounds of ethics and politics. They would then have to face the obvious: in terms of normativity, truth is on the side of folklore. The common intuition about our agency reverses the onus of proof: it's up to neuroscience to convince us that we don't have it.

Finally, we would like to present a few arguments in favor of resisting a potential neuro-conversion of criminal justice policies. We have already discussed the logical impossibility to go from positivity to normativity without additionally postulating that "what should be should conform to what is." This postulate, however, needs further elucidation.

Taken in a broad interpretation, this premise actually translates into a principle dear to justice: "no one is expected to do the impossible." To be fair, we can only ask of ourselves things that we can achieve. According to this principle, one might think that neuroscience is better suited to establish a basis for responsibility since, by definition, they would only require what

[9]This echoes a recent argument from Krakauer et al. (2017) in favor of *behaviourally* driven neuroscience: neuroscience needs behavior to make sense of neural findings. As a matter of fact, the *neural* implementation of behavior is always better investigated after having first carefully studied (i.e., theoretically and experimentally decomposed) the behavior itself (Krakauer et al., 2017; see also infra, "Technical limitations").

[10]*543 U.S. 551*, 2005.

[11]S.J. Morse, with humor, considers such arguments as "*the signs of a disorder that I have preliminarily entitled Brain Overclaim Syndrome*" (Morse, 2006, p. 397).

is accessible to human nature. However, this would be forgetting that law does not ask for perfection. To a certain extent, law is meant for the humans we are. When judging an individual's rationality, legal reasoning only takes common standards and expects an average fulfillment thereof. Neuroscience would thus not be fairer than law is in this regard.

The strict interpretation of the premise, i.e., the claim that description should translates into prescription (for example, taking the cortex myelination of teenagers as indicative of their lack of liability), weakens the law rather than consolidating it. Indexing normative standards to the current state of science dooms the latter to follow the vagaries of a branch of science that is necessarily evolving, often imperfect, sometimes flat out wrong while consensus arises and disputes settle. One notice in this respect the fast development of paradigms in cognitive sciences (from phrenology in the XIX century, to radical behaviorism in the 1930's, cognitivism in the 1950's, enactivism in the 1980's, etc.), and the legal incongruities that would arise from following such paradigms. This would lead to legal instability that goes against some fundamental principles of justice such as the necessity of having an explicitly enunciated law beforehand[12]. Past and current law, based partly on general criteria inspired from daily experiences, showcases continuity and stability, which science could not guarantee.

Moreover, the strict interpretation of the premise ignores a second principle dear to any normative framework, i.e., the principle of perfectibility. "Principle" might be too strong a term, and some might prefer using "aim." All things considered, perfectibility is a truism of normativity. A normative framework, while restrained to accessible requirements, still posits those requirements as desirable objectives to aim at. Those requirements can be mediocre, but everyone must at least aspire to mediocrity. The vision of a perfectible individual would be missing in a framework that ignores this aspiration. Such a framework would freeze men and women in their identified and limited abilities, without being able to legitimate the expectation that they give the best of themselves.

THE LIMITS OF NEUROSCIENCE

The previous paragraphs should not be read as ignoring the law's own flaws and limitations. Classical criticism of behavioral requirements for criminal responsibility points to the risk of circularity inherent to behavioral evidence, especially in assessing mental disorders: the absence of responsibility for antisocial acts would be assigned due to a mental disorder whose main, if not only, symptoms are those very same antisocial acts. That particular criticism has been amply discussed in the 20th century in a notorious debate opposing Lady Barbara Wootton and H.L.A. Hart (Matravers and Cocoru, 2014). Wootton supported that in *R v. Byrne*, "*the extent of Byrne's depravity was itself taken as evidence of his lack of responsibility*" (Wootton, 1963)[13]. While Hart nuanced that claim by reiterating the importance of circumstantial evidence at the time of the events in evaluating *mens rea,* the distinction between mad and bad remains a delicate one. In itself, a wrongful act does not sufficiently evidence the incapability of distinguishing between right and wrong, although the former is indeed a probable consequence of the latter. In the same vein, the more evil exceeds a reasonable person's imagination, the more it is associated with a deficient reason. Neuroscience might then be useful to the law. It could confirm or invalidate behavioral evidence. Besides, it already has been used in courts (see next section). However, precautions are once more in order. Neuroscientific evidence is restricted by technical and legal limits. We identify them here.

Technical Limitations

These limits are already addressed extensively in the literature (e.g., Pardo and Patterson, 2013; Kedia et al., 2017; Haushalter, 2018; Pardo, 2018). We will simply enumerate and describe them briefly:

Temporal Limitation

Neuroscience and its tools – especially brain imaging – can only prove permanent anomalies, still visible at trial, and not temporary conditions concurrent to the time of events and already dissipated at trial. Moreover, it is impossible to know whether the anomaly observed is anterior or posterior to the crime (Vincent, 2010, p. 95). Finally, as highlighted by others (Maibom, 2008; Reimer, 2008; Vincent, 2011), the permanent condition must also be linked to an inability to be responsible (i.e., an inability that paralyzes judgment) and not simply to a general feature of the accused's character, such as aggressiveness.

Interpretative Limitation

A first limit relates to the interpretation of functional imaging data (e.g., fMRI) and the risk of evidential circularity. Without diving deeply into the philosophical debate around mental states multiple realizability (e.g., Aizawa, 2009), it remains difficult to accurately map a cognitive process or function in a precise brain area, neural network or population. This difficulty arises from the fact that one brain area can perform different functions (*many-to-one* mapping) that are hardly distinguishable without an appropriate experimental protocol. Partially overlapping activity patterns associated with distinctive functions also complicates the proper interpretation of brain scans when they are not concurrently read with the patient's behavior (for example, when neural circuits required for an action's execution partially overlap with some linked to the observation of that same action executed

[12]The citizen needs to know the law so as to be able to comply with it, and this is made easier when the law is stable and does not follow the rather tumultuous course of scientific advances (see Hu et al., 2018). For example, and as illustrated by the recent Replicability Crisis, a variety of legally relevant notions in cognitive science (e.g., social priming, third-party punishment, biases in judicial decisions) might need to be profoundly revised, if not abandoned (e.g., respectively, Lakens, 2017; Schimmack et al., 2017; Pedersen et al., 2018).

[13]Byrne was a violent psychopath who mutilated, raped and killed a young woman in a youth hostel. The Court of Appeal defined the *abnormality of mind* as including the lack of ability to exercise will power to control physical acts in accordance with rational judgment. The Court held that the accused was in such an abnormal state of mind that he did not have the required *mens rea* for murder (the charge was reduced to manslaughter). However, the evidence of abnormality, according to Wootton, relied mainly "*the revolting circumstances of the killing and the subsequent mutilations* " as well as on "*his previous sexual history*" (Wootton, 1963). See R. v. Byrne (1960) *2 QB 396*.

by a third party, if not with the simple imagination of that action, see Jeannerod, 2001, for a review). The necessity of always going back to the behavior to interpret a functional scan makes brain activity evidence circular: it is used to prove or explain a behavior, and yet, brain activity patterns only mean something in so far as they are associated with the behavior they seek to explain (see infra, our criticism of P300-MERMER; see also Krakauer et al., 2017, emphasizing the better epistemological accuracy of *behaviourally driven* neuroscience). Hence, exclusive neural evidence, just as strictly behavioral evidence, does not solve Wootton's circularity issue mentioned above. Again, looking at the circumstances surrounding the alleged crime is necessary. Because brain scans are rarely informative in themselves – without referring to the behavior they seek to explain – there are few situations in which they are useful for establishing criminal liability. They may only be in distinguishing the truth in "gray area" cases "in which the behavioral evidence is unclear" (see Morse, 2019)[14].

A second linked limit is the risk of producing reverse inferences (see Poldrack, 2011), i.e., inferring a mental process from the observation of activity patterns without consideration for the actual behavior or the circumstances thereof. Reverse inferences can lead to fallacious interpretations of neuroimaging data such as: concluding that a blind woman sees because her visual cortex activates; or coming to the conclusion that dogs understand words of praise because some patterns, as revealed by fMRI, activate in their left brain hemisphere (Andics et al., 2016)[15]. It is worth noting that reverse inferences are often wrongly used as a common strategy to interpret experiment results. The problem is that neuroscience still does not have a sufficient understanding of brain functions to infer mental process on the sole basis of neural activity[16] (for a similar critic see Kedia et al., 2017). Reverse inferences, although tolerated in the context of exploratory scientific practices, is thus not fit for law's requirements, in particular considering the institution of criminal responsibility and the major consequences it brings about for an incriminated individual.

Let us note that this critique also targets the most recent tools used for probing neural activity, including brain data decoding techniques based on machine-learning (e.g., Multi-Voxel Pattern Analysis). The thesis that referring to behavior is essential to the correct interpretation of brain activity grows in importance when applying data-driven methods to decode the accused's intentions or thoughts. Indeed, nothing in "decoded" activity patterns alone indicates whether the brain is actually using those patterns to complete a task or to achieve a specific cognitive goal. In other words, it is still necessary to show that the pattern decoded by machine learning algorithms actually contributes to the studied behavior. This requires being able to explicitly link decoded patterns to behavioral outputs (e.g., Ritchie and Carlson, 2016). Without an explicit reference to behavior, decoded activity patterns have but weak explicative value: the possibility always remains that they might only reflect associative processes concomitant to the relevant functional process, e.g., the reuse of sensory information for higher-level operations (Ritchie et al., 2017; Bouton et al., 2018, for a review).[17]

Comparative Limitation

To be significant, fMRI scan results must be replicable and subjected to group analysis. An fMRI scan is a functional scan that measures and maps brain's activity while the subject is completing a task (e.g., encoding information, storing it, using it to make or guide decisions, etc.). Specifically, what is measured is an indirect effect of brain activity, i.e., a modification of oxygen levels in local blood supplies (blood-oxygen-level-dependent response, or BOLD signal). This measurement is considered as a reliable indicium of a specific brain area being required to do a task, if not essentially "doing" that task. However, linking BOLD signal variations to cognitive processes remains difficult for three reasons: (1) even in a resting state, the brain presents spontaneous activity fluctuations; (2) neural computations have intrinsic noise; (3) what one does or what one thinks in a scan can never be completely controlled. It is thus imperative, before introducing fMRI scans in courtrooms, to conceive experiments carefully designed to isolate, in an individual's brain, activity fluctuations relevant to the behavior being studied, i.e., experiments (factorial or parametric designs) that discriminate between relevant neural activity and background or task-unrelated neural activity.

In this regard, Kedia et al. (2017) recall the importance of replication and generalization in order to assess fMRI measurement reliability. These require a great number of observations/acquisitions in the view of minimizing the signal-noise ratio, as well as replicating results between individuals or cohort in order to avoid statistical artifacts. Accordingly, the interpretation of functional scans from a single person (for example, the accused in a trial) is extremely dubious as it is vulnerable to type I (false positive) and II (false negative) statistical errors that can only be avoided through robust group analysis and rigorous experimental protocols.

[14] Morse (2019), in fact, seems less optimistic than we are: "*if a criminal defendant behaves rationally in a wide variety of circumstances, the defendant is rational even if his or her brain appears structurally or functionally abnormal. In contrast, if the defendant is clearly psychotic, then a potentially legally relevant rationality problem exists even if his brain looks normal. We might think that neuroscience would be especially helpful in distinguishing the truth in "gray area" cases in which the behavioral evidence is unclear. For example, is the defendant simply very grandiose or actually delusional? But unfortunately, the neuroscience helps us least when we need it the most, and if the behavior is clear, we don't need it at all*".

[15] A number of articles have interpreted this result as signifying that dogs *understand* human words because lexical processing is associated with a similar pattern of activation in the left hemisphere in most humans (but see also Andics et al., 2017, *Erratum for the Report "Neural mechanisms for lexical processing in dogs"*).

[16] Among other examples, there are inconsistencies in brain areas associated with moral reasoning: utilitarian decisions (sacrificing one life to save three others) in the Trolley dilemma recruits a structure located in medial part of the prefrontal cortex (the anterior cingulate cortex), while it has been shown that damage to prefrontal regions increases the frequency of utilitarian decisions (Capestany and Harris, 2014).

[17] The fact that in linear classification (the method used by most decoding techniques) there is little constraint on how information is selected and classified is both the strength and weakness of the technique. This explains why classifiers can robustly decode features in brain regions that are yet known to code poorly for these features (e.g., visual motion in V1, Seymour et al., 2009; Wang et al., 2014) or can decode arbitrary univariate fMRI signals that classical activation-based analyses could not detect (e.g., Davis and Poldrack, 2013).

Normative Limitation

The relevance of results, be they from functional or anatomical scans, depends on the (normative) definition of handicap linked to a certain behavior. For example, anatomical scans (the equivalent of pictures of the brain structure) can reveal anatomical alterations and anomalies (e.g., loss of cerebral matter, alteration in the organic structure, excessive spinal fluid, etc.). Relevantly producing such evidence, however, implies the hypothesis that those anomalies alter the accused's capacity to follow or detect a norm, or to adapt to or adopt an appropriate behavior. Anatomical anomalies alone do not indicate the presence of a handicap, and do not necessarily translate into mental deficiencies. Extreme examples exist of people having one entire hemisphere removed (hemispherectomy) and yet, not experiencing any abnormal difficulty in their daily lives, even when the hemispherectomy has been performed at a late stage of development (Schmeiser et al., 2017)[18].

A functional or anatomical anomaly is interpreted as being a handicap only insofar as the behavior it might produce is considered such. To say that a subject is not able to follow the rules due to brain injuries requires proving that these injuries are the source of that disability (as indeed, most penal codes prescribe). Neuroscientific tools may thus indicate the source of a disability (and not be the evidence of the disability itself). Yet, although some scientific findings prove that some prefrontal injuries generate sociopathic tendencies (e.g., Phineas Gage), not all prefrontal lesions lead to such tendencies. Structure–function mapping is, in fact, relatively flexible. Further, the brain is functionally vicarious: under certain conditions, new functions can emerge via the reuse, the recycling, or the reconfiguration of existing brain circuitry (e.g., Anderson, 2010; Wittenberg, 2010). Interpreting functional or anatomical anomalies remains questionable, and cannot forgo referring to the abnormal subject's behavior.

Experimental Limitation

Laboratory conditions and actions typically tested do not necessarily reflect the conditions of daily life in which individuals normally act (see **Box 3**). Participants' movements, for example, are extremely restricted in a scanner (any head movement superior to a few millimeters can jeopardize results and produce false positives[19]). Experiments testing an agent's intention, choice and responsibility are more exposed to this line of criticism. Some have argued that the actions participants are asked to perform (such as pressing on buttons or targets, or following a sequence of buttons pressed following an audio signal, etc.) are not intentional since they are not chosen. More precisely, they are triggered by exterior conditions/demands and they are almost automatic, without any surprise and spontaneity as to the when and how (Brass and Haggard, 2008; Waller, 2012). Furthermore, A. R. Mele shows that what is called "intentional" varies from scientists to philosophers, and that some actions can be considered as intentional even when following strict instructions or when not being fully conscious (Mele, 2009; Chambon et al., 2011; see also Pacherie, 2008, for a three-tiered dynamic model of intention). Despite this nuance, it is obvious that "*the arbitrary free choice afforded participants in the experiments, the choice of when or whether to perform a simple movement, is disconnected from participants' everyday justificatory or motivational reasons—moral, prudential, or otherwise—for action and thus fails to capture the type of decisions and actions for which agents are typically held morally responsible*" (Waller, 2012)[20]. Neuroscience could nonetheless compensate for this shortcoming through revisited protocols.

As a final remark, it is worth pointing out that neuroimaging can (and will undoubtedly) contribute to make the assessments of criminal liability more objective than other – and sometimes more idiosyncratic – behavioral assessment tools within the traditional context of criminal law. While saying this, we must also recall that law's criteria are first and foremost behavioral – actions and mental states are what are judged. Thus, while we recognize that classical behavioral assessments can be distorted by the expert's subjectivity, it should also be noted that behavioral data can readily be translated into notions that speak the language of law, while neural data are rarely self-explanatory, especially not with respect to the defendant's behavior (see above for the evidential circularity of functional neuroimaging evidence).

Legal Limitations

Legal limitations might be more severe than technical limitations since overcoming them depends on exclusively legal debates. However, they inform neuroscientists who wish to assist the courts or to simply legally contextualize their scientific findings.

First, neuroscience can only impact legal excuses and not legal justifications. By definition, the latter concerns external restrictions to an agent's actions. An agent's actions will be justified due to the existence of only one reasonable solution to a problematic situation. Arguments relating to neurological conditions reducing possible options (such as "my brain was in such a state that it was impossible to avoid acting a particular way" or "my brain did it, not I") do not intervene at this stage. Justifications do not only tackle phenomena out of will power's reach (like electrical pulses in neural circuits), but precisely phenomena completely independent and external to the agent, including its neural circuits. Justifications are about circumstances external to oneself, or even actually contrary to oneself since all the goodwill in the world could not prevent wrongdoing. This is the case with self-defense, for example, when

[18] See also Nahm et al. (2017): "*Large amounts of brain mass and its organic structures, even entire hemispheres, can be drastically altered, damaged, or even absent without causing a substantial impairment of the mental capacities of the affected persons*". About a patient with hemispherectomy, "*not only does [the patient] perform motor and sensory functions for both sides of the body, [he] performs the associative and intellectual functions normally allocated to two hemispheres*" (Nahm et al., 2017).

[19] A non-consensual participant needs only move his or her head slightly to render the results uninterpretable. Thus, it has been consistently shown that subject motion in fMRI produces spurious but systematic correlations in functional connectivity, which are interpreted as true correlations while they are in fact simple motion artifacts (e.g., Power et al., 2012; Van Dijk et al., 2012). Note that the same remark applies with twitches, blinks and fidgets, as important generators of ongoing neural activity (Drew et al., 2018).

[20] "*Given that the types of actions at issue in the free will and moral responsibility literature are often preceded by deliberation and are actions according to which we evaluate the agent, the lack of these features in the experiment might seem unsatisfactory.*"

BOX 3 | Representativeness of fMRI participants.

The representativeness of fMRI participants has been questioned. For example, people who do not meet inclusion criteria for fMRI scanning are automatically excluded from neuroimaging studies, including individuals wearing tattoos or permanent jewelry, devices or metal in their body (whether aneurysm clip, pacemaker, or metal fragments), pregnant women, etc. Also, most neuroimaging data are collected from student subjects pool, and from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations more broadly (on WEIRD people, see Henrich et al., 2010; see also Baumard and Sperber, 2010 on WEIRD experiments). These samples may differ in many concrete ways from broader populations of interest (Falk et al., 2013). Early life experiences are rarely taken into account when screening and recruiting participants; yet parenting and socio-economic status (SES) have effects on brain areas such as the amygdala and prefrontal cortex, whose dysfunction has been linked to a variety of legally relevant outcomes such as crime and violence, drug use, and reduced cognitive control (see Falk et al., 2013, for a review)

Statistical reliability of fMRI results

The reliability of neuroimaging results has been the subject of much discussion (for a review, see Eklund et al., 2018). Various software used in fMRI analysis have bugs that increase the rate of false positives, i.e., the probability of finding a significant activation (yet a statistical artifact) in a specific region during a given task. In a recent paper, Eklund and colleagues estimated that about 10% of the fMRI experiments in the literature – thousands of fMRI studies – were in doubt and could have produced at least one false positive. It is possible to control the false-positive rate in fMRI by correcting from multiple comparison, a gold standard of statistical massively univariate analyses such as fMRI. However, the type of correction that should be used is also a matter of discussion (e.g., Woo et al., 2014). Indeed, an appropriate balance must be found between trying to minimize false positives (Type I error) while not being too stringent and omitting true effects (Type II error) (Han and Glenn, 2018).

Ecological validity of fMRI experiments

Serious doubts have been raised as to the admissibility of fMRI evidence in judicial settings. Due to their lack of ecological validity, neuroimaging studies – laboratory experiments in general – can prompt behaviors that have no real functional meaning but in the constrained space of the scanner. This could be the case of the so-called "altruistic punishment" behavior, whereby individuals "punish" defectors or free-riders although the punishment is costly for them and yields no material gain (Fehr and Gächter, 2002). However, what is observed in natural settings draws a different picture: individuals who are identified as free-riders are generally not "punished" but either ignored or simply excluded from any subsequent transactions in favor of other, and potentially fairer, partners. In other terms, laboratory volunteers would engage in altruistic punishment because, in the reduced space of the experimental room, they would not be given "outside options," e.g., the opportunity to find more cooperative partners (Guala, 2012; see also Barclay and Raihani, 2016). This observation echoes a recent study showing that people punish altruistically because the experimental setup (an economic game with oriented instruction) incites them to do so – a phenomenon known as "experimental demand" (Pedersen et al., 2018).

circumstances someone faces only allow for two options - kill or be killed-, knowing that the latter option constitutes the threshold beyond which obedience becomes illegitimate.

"Impossibility" is also a corresponding line of defense. Its definition in Canadian law is precisely "an exterior, unpredictable and irresistible cause that prevents the individual, despite his or her own will, from conforming to the law"(Parent, 2008, p. 769)[21]. "Necessity" is another legal justification that follows the same rationale, although more flexible as it allows the possibility of choosing between two evils. Aristotle notoriously illustrated the situation of a mixed act (intentional but constrained) through the story of a captain's ship.[22] Moreover the standard of appreciation of all those justificatory factors is objective, which means that it applies the standard of "the reasonable person" placed under the same circumstances (Parent, 2008). Objective evaluation in these cases serves the purpose of knowing whether or not the alleged crime was bound to happen independently from the accused's personal characteristics. In this regard, scientists should pay particular attention not to comment on legal justifications when addressing the issue of criminal responsibility[23].

Finally, any evidence submitted at trial, be it scientific or not, has to be validated by certain legal tests before being accepted and presented to a jury. These tests generally ensure that the accused's rights and the constitutionality of investigating methods are respected. They allow, for example, excluding evidence (even if overwhelming) that would come from an unlawful search in the accused's house. A similar degree of vigilance applies to technical evidence, such as expert testimony, medical reports, etc. In American law, for example, evidence must be admissible and relevant.[24] One of the criteria for admissibility, as elaborated in Frye v. United States (1923) and known as the Frye Test, is the general recognition of the evidence's experimental value by the appropriate scientific community (see **Box 3**). While adopted just under a century ago, the Fye Test still serves today to exclude non-consensual techniques, e.g., to restrict the use of genetic evidence of behaviors in federal habeas corpus cases (Cullen v. Pinholster, 2011; Kaufmann, 2013). The Daubert trilogy in 2002 then clarified and modified provision 702 ruling over testimonies and expert reports[25]. The Daubert Test establishes the following admissibility conditions: (1) the expert report must be based on sufficient facts and data; (2) the testimony is based on reliable principles and methods; and (3) those

[21] For example, driving carefully, and yet above the speed limit, when a snowstorm prevents the driver from seeing the road signs.

[22] The act is intentional, but constrained. This type of excuse acknowledges the presence of *mens rea*: in *the Nicomachean Ethics*, Aristotle illustrates the situation of a mixed act by using the image of a captain's ship in a storm who must abandon his shipment to save his crew. In this case, the captain's action results from the captain's choice, and hence it is still a voluntary action even though the action was constrained by external causes.

[23] For example, the notion of self-defense is sometimes used to illustrate a claim about responsibility, including in cautious and relevant articles (e.g., Haggard, 2017).

[24] Rule 104 Fed. R. Ev. : "*104 (a) Preliminary Questions of Admissibility, and (b) Relevancy Conditioned on Fact, as follows: (a) The court must decide any preliminary question about whether a witness is qualified, a privilege exists, or evidence is admissible. In so deciding, the court is not bound by evidence rules, except those on privilege; (b) When the relevance of evidence depends on whether a fact exists, proof must be introduced sufficient to support a finding that the fact does exist. The court may admit the proposed evidence on the condition that the proof be introduced later."*

[25] Rule 702, Fed. R. Ev.; [*Daubert v. Merrell Dow Pharmaceuticals, Inc., 43 F 1311 (9th Cir. 1995)*, s. d.]

principles and methods have been faithfully applied to the facts in question. Those criteria are, however, neither exhaustive nor exclusive, and others have been developed: whether the evidence submitted belongs to the expert's usual field of research or on the contrary have been elaborated in anticipation of the trial (Daubert v. Merrell Dow Pharmaceuticals, Inc, 1995) the consideration of alternative interpretations, (Claar v. Burlington, 1994) the influence a lucrative contract might have exercised over the expert's diligence (Sheehan, 1997), the general reliability of the expert's field of study Kumho Tire Co. v. Carmichael (1999) the presence of extrapolations in the expert's reasoning, General Elec, Co. v. Joiner (1997) and others.[26] The more recent case of Terry Harrington v. State (2003) set even clearer and more concise admissibility criteria: (1) the previous publication of the submitted tests and methods in blind-peer reviewed journals, (2) the testing of these methods outside laboratory in real life conditions, and (3) scientific community's approval thereof (Frye Test) (Pallarés-Dominguez and Esteban, 2016) (see **Box 3**). Besides being admissible, evidence must also be relevant pursuant to provision 403 Federal Rules of Evidence. It must be highlighted as well that these tests, although similarly and generally requiring admissibility and relevance, vary from one jurisdiction, country and legal tradition to another.[27]

Now that leeway for neuroscience has been defined, we can look into concrete attempts at introducing such techniques into courtrooms.

Lie Detectors

A P-300 MERMER test (Memory and Encoding Related Multifaceted Electroencephalographic Response) or *Dr. Farwell's brain fingerprinting* (e.g., Farwell and Smith, 2001) is not exactly a lie detector. Rather, it highlights the accused's memory, or absence thereof, about certain facts, by measuring a positive brain wave called P300 MERMER. A certain wave potential obtained through relevant stimulus would show the presence of an actual memory linked to this stimulus. Proponents of this technique measure the wave amplitude from P300 responses to images or words linked to familiar events or events recognized by the accused: a crime, terrorist training, bomb-crafting knowledge, etc. The test produces a neural signature for the absence or presence of relevant information in the accused's memory, and gives a reliability index for that result. Experiments in and outside the laboratory have shown an error ratio of less than 1% (Farwell, 2012, for a review). P-300 MERMER test has been used in a somewhat contradictory manner in the courts: in Harrington v. State (2003) it allowed for the release of a man wrongly convicted of murder after 23 years of imprisonment. However, in *State v. Grinder,* it has been recognized as a highly probative and incriminating evidence (Harrington v. State, 2003; Brandom, 2015). Other techniques have been developed, such as a TMS (transcranial magnetic stimulation) procedure that disrupts brain areas supposedly implicated in intentional trickery (e.g., George et al., 2006; Rosen, 2007), but they present less reliable results.

Those methods are questionable on many levels. Conceptually speaking, they contribute to "mereological fallacy," that is the general tendency of neuroscience to ascribe to the brain, or parts thereof, abilities or properties that in fact belong to individuals. It wrongfully attributes a property of the whole to one particular mechanism (Pardo and Patterson, 2013). However, this conceptual objection is not consensual (Levy, 2014). In the same vein, the possibility of detecting lies is contested by the mere context-dependant definition of lying: "*As Don Fallis notes in an insightful article, the difference that makes "I am the Prince of Denmark" a lie when told at a dinner party but not a lie when told on stage at a play are the norms of conversation in effect*" (Pardo and Patterson, 2013). A false declaration is thus not always a lie and depends on whether or not it is stated in a conversational context whose norm is "you shall not make false declarations." Nevertheless, participants in lie detecting experiments are precisely instructed to utter false declarations, and therefore perform in a context antithetical to lying. Besides, someone can lie without knowing it, when stating something false and yet believing it is true (Faulkner, 2007). One can also convince oneself that a false information is in fact true (Van Horne, 1981; see also Pardo, 2018, for a critical review). In other words, what neuroscientific tools record are not lies. On a more practical note, some authors worry that it would already be possible to elaborate counter-measures in order to cheat lie detectors (Kedia et al., 2017).

On a strictly technical level, P-300 MERMER test results are more than doubtful. Most of the studies on which that method rely, focus on small biased samples (often student volunteers rather than real accused in real investigation conditions). Most of the studies on the accuracy of that method are rarely reviewed on a blind-peer basis.[28] Moreover, the 20 fingerprint standards defined by Farewell himself to evaluate his own method's efficiency are controversial. Some scientists consider them to be purely subjective and self-confirmative, as they are not defined by a scientific consensus (Meijer et al., 2013). One of the most severe criticisms comes from one of Farewell's mentor (Dr. Donchin), who criticized the (laboratory) conditions in which it has been mainly tested. In real life conditions, some parameters must still be addressed: the reliability and efficiency of the electrophysiological response for real accused persons, for example, or neurologically atypical individuals. In sum, given the large differences between the typical experimental setting and realistic criminal investigations, it is questionable whether the results of P300 MERMER experiments can be generalized.

The relative impossibility of replicating Farewell's method for independent researchers, at least with similar statistical power, should also be noted. Indeed when replications take place, results show less statistical strength compared to the original studies (88% of correct detections (Meijer et al., 2014), which is similar to results obtained through other techniques linked to the autonomous nervous system (Skin Conductance Response, Respiration Line length, Changes in

[26] All information from this paragraph comes from an excellent procedural review on the question provided by Kaufmann (2013).

[27] For a detailed comparative law study, see Spranger (2012).

[28] According to Meijer et al. (2013), in the seminal line of research from Farewell and collaborators, only two studies were peer-reviewed – that is to say, 3 datasets with a total of 30 participants (i.e., Farwell and Donchin, 1991; Farwell and Smith, 2001).

heart rate, etc.). P300 MERMER's accuracy is also vulnerable to counter-measures (see Rosenfeld, 2005, for a comprehensive review). It actually collapses when crime-based items are compared to irrelevant items with the largest P300 responses (Lukács et al., 2016). Finally, the test lacks sufficiently reliable baseline measures, that is: truly neutral questions asked to participants.

Mnemonic recognition of familiar details of events is at the heart of P300 test. This mnemonic recognition is nevertheless challenged by other limitations: the fact of having crime-based information stored in memory is not sufficient to infer guilt, as frequent or significant details for a participant might trigger that very same event-related potential; P300 is susceptible to false memories[29] and also to lack of participant attention ("many guilty suspects ended up passing the test simply because they hadn't paid attention to the objects in the test," see Meijer et al., 2014). Finally, some have gone to the extent of questioning the whole of P300's relevance and argued that the test's benefit lies only in the examination strategy used (Classification Concealed Information, CIT) and not in the electrophysiological signal itself.[30]

One last legal objection is possible. P300 MERMER might indeed violate the right against self-incrimination.[31] This right is one of the fundamental rights of the accused, namely the right to silence, the presumption of innocence (which shifts the onus to the Prosecution to prove allegations beyond reasonable doubt) and right to not be compelled to give evidence at one's own trial, etc. In the case of Antonio Losilla, the argument was raised in court to appeal the decision authorizing P-300 MERMER (Lukács et al., 2016). Nevertheless, the test had been used before the decision was rendered. In the United States, *Schmerber v. California* found that the 5th Amendment protected the accused from being forced into "*prov[ing] a charge from his own mouth*" but that it did not apply to material and physical evidence. That distinction between verbal and physical testimony has since been roundly criticized by jurists for its inconsistency with the objective of the right against self-incrimination (Farahany, 2012).

The main objections are conceptual, technical, and legal, and although each is limited in its own scope, together they seriously bring into question the use and rigor of such methods.

[29]The P300 component reflects the subject's *beliefs* rather than the recognition of real facts – but even false memories can return positive results (Satel and Lilienfeld, 2013).

[30]This point echoes the critique raised about the use of fMRI scans in judicial settings, and the risk of evidential circularity (see supra, "Interpretative limitations"). As pointed out by a influential neuroscience blogger: "*What do we do about someone whose brain "lights up" to the taboo stimuli (child, or pro-terror), but who denies feeling any attraction? What about someone who acknowledges a taboo desire, but who has never acted upon it and who says they never will? Neuroscience might offer a source of information, but we'd still have to make sense of that data,*" i.e., to refer to the actual behavior (Neuroskeptic, "Do We Need A Neuroscience of Terrorism?", Discover magazine). A similar remark is made by Coppola (2018): "*There can be cases in which individuals who experience paedophilic urges [and display neurobiological profile associated with paedophilic traits] are still able to resist them.*"

[31]This right applies in the United States, Canada, Wales, England, and India.

NO REVOLUTION, ONLY DIALOGUES

Neuroscience's claims relating to law generally can be separated into three categories: (i) revision or reform, according to which neuroscience overthrows current legal criminal standards; (ii) *evaluation*, which consists of using neuroscientific tools to play a role in the judicial process; and (iii) *intervention*, which translates into the direct manipulation of people's brains (this clever classification is borrowed to Meynen, 2014). We have already established through Section "What Is Criminal Responsibility?" and Section "Understanding Normativity: Neuroscience Tells but Does Not Compel" that revisionist claims have no foundations. While keeping in mind the limitations addressed in Section "The Limits of Neuroscience," we would now like to focus on cases suggested by the other two remaining categories, and will deal with several attempts at introducing neuroscientific elements in courtrooms.

Irresistible Urges and Rationalism

Criminal law generally adopts an intellectualist/rationalist approach (as opposed to volitionist/will oriented approach) in evaluating an agent's capacities. That is, it seeks to determine whether or not an accused has a functioning sense of reason, and not to assess the strength of his or her will. It thus recognizes deficiencies of rationality but not weakness of will. In Canadian law, for example, provocation is a defense that reduces murder to involuntary homicide due to a violent anger provoked by "an action or an insult of such a nature as to be sufficient to deprive an ordinary person of the power of self-control"[32]. It relies on the evidence of a momentary lapse in judgment and not a simple urge (see also **Box 4**). The expression "self-control" (and loss thereof) are not controversial and are associated with a "temporary suspension of reason" or "*the temporary eclipse of reason by passion as the guiding force influencing one's action*"(R c. Gibson, 2001). The same rationalist approach applies to other behavioral disorders, such as pyromania and kleptomania. Being a kleptomaniac is not sufficient grounds for being exonerated from stealing because criminal law considers that a kleptomaniac still knows that what he or she is doing and that stealing is wrong. Some debates still shake the legal and philosophical community as to the validity of pleading irresistible urges and the voluntary aspect of acts, but rationalism prevails (Morse, 2002; Parent, 2008, p. 859).

Neuroscience would here claim that some behaviors that we take to be malevolent urges are in fact deficiencies of reason.

One of these claims relate to drug addiction. Neil Levy hence contends that drug addiction is not to be considered a compulsive behavior but rather to as altering judgment capacity: "*though most of the time addicts judge that they ought to refrain, at the time of consumption they judge that all things considered they ought to consume*" (Levy, 2014). This alleged contradiction would show that drug addicts suffer not only from a shortcoming of will power but also a disorder of reasoning. Moreover, drug addicts' endorsement of their own behavior is equivocal. Neuroscience would then be more suited than behavioral evidence to establish

[32]Art. 232(2) C.cr. (Canada).

BOX 4 | Criminal responsibility under influence.

Some concerns accompany the growing use of invasive technologies such as neural implants and "deep brain stimulation" (DBS) neurosurgical procedures. Patients having received DBS as treatment may exhibit various side effects, from developing new musical preferences to suffering from temporary hallucinations. What about cases where the implant would be the cause of a criminally blameworthy behaviour? Do these brain devices entail a revision of our legal categories about responsibility, just like assisted reproductive techniques have changed the legal definition of a parent?

Once again, the current law already has tools to address situations of potential concern caused by DBS. If the accused at the time of events perceives the reality differently from what it is (hallucination), she/he cannot be held responsible. The expert evidence about the role of the implant in the false perception would not be in and of itself exculpatory, but it would add to the credibility of the defence's narrative. More generally, the law recognizes *intoxication* as a defence, as a state or an external influence that alters the accused's perception and personality. *Voluntary intoxication* – drug and alcohol abuse – is not exculpatory because it is assumed that the accused knew about the adverse effects of the substance beforehand.

Involuntary intoxication (which may correspond to the side effects of a medication) is recognized as a valid defence. Then the question would be to know whether the potential adverse effects of DBS could be assimilated to involuntary intoxication. Thus, it would be possible to modify, and even rename, an already existing defence – involuntary intoxication – to include new interfering influences. However, this "new" defence would follow the same logic as the previous one. Once again, brain technologies do not revolutionize law but improve and marginally modify existing legal resources (see Klaming and Haselager, 2013).

In a similar vein, liability issues concerning the releasing of risky technology into the market are not a novelty. Our case could be compared to the pacemaker in this regard. The law already addresses many aspects of this issue (the patient's consent, knowledge of the risks, transparency concerning the risks, professional insurance for doctors, etc.). We wish not to speculate about the wording of future provisions to deal with DBS but rather to stress that the law is already well equipped to deal with seemingly new objects.

that link, and could as a result lead to a not-guilty verdict or a verdict reflecting a diminished degree of responsibility.[33]

However, we can object that numerous drug addicts report knowing that what they do is wrong. They do not showcase a troubled reason that would not dissociate right from wrong. Levy argues that it is possible to be wrong about one's own mental state and that subjective experiences can thus be erroneous (similarly to cases of erroneous affective attribution or cognitive dissonance). We cannot admit this answer: the flaw in a subjective experience is relative to a context and to an external observer, not to a neurological state. In other words, a drug addict who is acting illegally while cognizant that he or she is acting as such has no rational deficiencies. An external observer can only but note that action, thought and reality are all in agreement. We broadly consider that a person claiming to see Satan is mad because this subjective experience does not correspond to reality (again, the normative reality of a secular law that does not acknowledge Satan's existence). What is deemed a bad judgment is normatively qualified from the outside. Cognitive disorders are disorders for the experts observing them. The subjectivist objection that calls for considering the subjective experience of drug addicts is thus valid.

The case of drug addicts reporting thinking that, at the moment they act, they are acting as they should, still needs to be addressed. Levy's argument here takes advantage of the ambiguity of terms like "should/right/duty." If science can show that drug addicts think that they are doing the right thing or accomplishing their duty while committing crimes, they would indeed demonstrate the delirious nature of drug addiction, and thus the judgment deficiencies it brings about. Levy, relying on Yaffe (2013), nonetheless seems to adopt a more personal definition of "duty" and confuses it with "value." Drug addicts would not think that they are accomplishing an objectively (normatively) good action, but a good action according to their own values.[34] Yaffe claims that there is a legal difference between a behavior guided by the agent's own values and, conversely, one that goes against them. Asking such drug addicts to respect the law is to make them bear too heavy of a burden. Accordingly, they should be findings of diminished responsibility should be available to them.

Again, we doubt the validity of such arguments due to prevailing normative standards. Criminal law currently judges even more harshly people who respect their own values at the expense of respecting the law. Let us recall honor based crimes as examples, or the very definition of misconduct ("*faute*" in French) for that matter. More precisely, let us take the example of provocation in Canada: an accused will be able to argue that the insulting attitude of a soon-to-be ex wife's new lover amounts to provocation, but will not be able to do the same about a homosexual flirtation, even where the accused is homophobic.[35] It is indeed hard to obey laws we don't value. We are nonetheless responsible for disregarding our values to the benefit of those laws.

Despite the weakness of some of Levy's arguments, it is worth noting the interesting idea they bring about, namely the possibility of clarifying some compulsive disorders and "neurologising" psychiatry (which means to seek to describe psychiatric disorders in terms of organic deficiencies, or on the contrary, establishing psychiatric diagnosis only once the organic causes are excluded). We don't exclude the possibility that neuroscience could one day demonstrate that drug addiction, or even pedophilia, translates into judgment disorders. They will then have to establish this while keeping in mind the rationalist criteria of criminal law (relation to common reality, ability to distinguish right from wrong, etc.) and addressing typical criticism concerning compulsive behavior, e.g., blame that rises from the fact that no measures were taken by the accused to avoid wrongdoing, even though he or she knew about his or her condition (a kleptomaniac could warn the shop owner, the drug

[33]Note that the defense of intoxication can be raised for crimes of specific intent (e.g., murder).

[34]Indeed, various studies have shown that pathological gambling is associated with a specific pattern of subjective preferences, characterized by a shift toward risky options (e.g., Ligneul et al., 2013). It should be noted that the interpretation of the word "duty" made by Yaffe derives from a distortion of the common word "value". A preference for risk-seeking strategies is not axiological.

[35]Respectively: *R c. Thibert,* [1996] 1 R.C.S. 37, 52, and *R. c. Tomlinson,* [1998] S.J. (Quicklaw) n848 (Q.B.).

addict could ask for help, a pedophile could avoid working in kindergartens, etc.).

Cognitive Biases

The "reasonable person" standard is often used in criminal law when objectively assessing the accused's *mens rea*. It generally serves in cases of omissions rather than actions[36], since for the former it is harder to evaluate the presence of a clear intention. Indeed, some acts speak for themselves, and we can almost intuitively guess the intention behind them. However, for others, when the accused actually did "nothing" and let the events occur, it is hard to positively find an intention. To know whether or not an attitude is wrongful, we then imagine "a reasonable person" facing similar circumstances. For example, leaving a toddler to play alongside a staircase could be considered criminal negligence, since any reasonable person could foresee that this is obviously not a good idea that will, in all odds, result in a tragedy.

Some studies reveal daily cognitive biases and suggest that the "reasonable person" standard be amended by such findings. Those studies outline, for example, a natural inclination to be overly confident in one's own judgments (overconfidence effect, Pallier et al., 2002), to filter information confirming these judgments (confirmation bias, Nickerson, 1998) and to ignore or discard conflicting information (bias against disconfirmatory evidence, Buchy et al., 2007); or even the natural tendency to believe that our successes are our own but that our failures are due to others or to external circumstances (self-serving bias, Shepperd et al., 2008), etc. Otherwise put, the reasonable person might not be that reasonable according to classical standards of rationality (e.g., Gigerenzer and Goldstein, 1996).

This is why Dahan-Katz (2013) criticized the judicial decision in Keech v. Commonwealth (1989). In this case, a driver was driving on the wrong side of a highway, while still believing he was on the right side. He persevered for 8 miles without understanding or paying attention to the other drivers' warnings, and finally caused a deadly accident. The tribunal found him guilty of manslaughter (different from murder) based on the fact that he should have known that he was driving dangerously. Dahan-Katz nonetheless argues that it is plausible that Keech was influenced by a bias according to which "*where a person is under the impression that a hypothesis is correct, indications to the contrary are not necessarily "rationally" considered—beliefs tend to persevere more than they ought to*". He should therefore have been relieved of all charges.[37]

This suggestion, although stimulating, seems to ignore that the "reasonable person" standard does not call for perfection. It does not refer to the perfect citizen but to the average person. The accused is not required to have rationally taken into consideration every aspect of the situation, but is rather asked to have considered it as an average person would have. However, severe our biases, and regardless of their effect on our rationality, we all share the same and it is according to this norm that we judge each other. We may indeed have a tendency to overestimate our abilities, but Keech's strange case nonetheless points to an all but ordinary behavior.

Cognitive neuroscience's claims in this regard could be more nuanced: it wouldn't inform the law about human frailty (which the law already takes into account) but would weigh in favour of a change of paradigm, from classical rationality standards (even if mediocre, degraded, or bounded; see Gigerenzer and Selten, 2002), to adaptive rationality criteria (Haselton et al., 2009). Persisting in believing one is right when one is wrong, for example, is considered irrational from a classical standpoint, and yet, is completely legitimate on an evolutionary level in terms of fitness (or cost rationale indicating that it costs more to change for an uncertain benefit than to persist in error) (e.g., Haselton and Nettle, 2006).

It can first be re-affirmed that the classical requirement for banality already acknowledges human biases and weaknesses (all the more so since cognitive psychology deals with biases that we experience on a daily basis). First and foremost, adaptive rationality cannot account for the principle of perfectibility present in and necessary to criminal justice. Classic rationality is referred to in the law as reasonableness in order to be accessible to the average citizen. Under this appellation, it retains the mark of an ideal to strive for, and still asks of people that they do their best to achieve that ideal. Adaptive, or bounded rationality is indifferent to the principle of perfectibility. Concretely, it indicates biases' functions, but it cannot demand to correct them, since those biases can be viewed as "adaptations." It would only require from people what they minimally already are (and it could certainly not prognosticate on biases adapted to the future). Only the classical ideal of rationality, inherent to its ideal nature, can call for more. Some may consider it as out-dated or excessively onerous. Yet again, law requires only an average rationality, a degree of reasonableness that is relative to a historical, cultural and punctual context. In doing so, it does not abandon the idea that it is *right* or *good* for humans to strive to respect the law by virtue of their capacities and choices. Cognitive neuroscience, and related disciplines (e.g., cognitive psychology, neuroeconomics) would thus not change (or should not change, depending on our ideological attachment to the principle of perfectibility) the paradigm of the "reasonable person" standard, but would *inform* this paradigm with the objective of providing a scientific basis for understanding what standard of reasonableness a particular person might be held to.

Nonetheless, cognitive biases indicate other avenues than the revision of the reasonable person standard, such as training for judges and juries. These could be useful to warn the latter about potential biases in their judgment and that of others. A famous, but controversial, example is a study supposedly showing that judges render harsher decisions when they're hungry (Danziger et al., 2011; for critics, see Weinshall-Margel and Shapard, 2011; Lakens, 2017). Another classical example comes from

[36]The standard of a reasonable person does not only apply to omissions, it also applies to many active offenses, such as the reasonable foreseeability requirement in aggravated assault, or the standard of care of the prudent driver in dangerous driving. We overly simplify its application to give the reader a grasp of what this standard is aiming at.

[37]Here Katz overstates the explanatory scope of cognitive biases in general: a confirmation bias can explain why an individual perseverates in performing an erroneous behavior, but it does not explain why this behavior has been adopted in the first place, e.g., it cannot account for Keech's initial decision to drive on the wrong side of the road (at the most it could explain why it lasted this long, but see below for a counter-argument).

studies on eyewitness testimony. Memory of an event that has been witnessed is highly flexible. Exposing a witness to new information during the interval between witnessing the event and recalling it, can substantially modify what the witness recalls (Loftus and Palmer, 1974). Evaluation of eyewitness evidence should probably be more attentive to this issue. Testimony, although essential and relevant evidence, could be considered less reliable, or at least elevate the burden of proof. Again, cognitive psychology findings provide useful tools, but do not radically transform legal practice. Lawyers, before the rise of psychological evidence about the frailty of our judgment and perceptions, have always proceeded in questioning witness and testimonies' credibility.

Sentences and Damages

It has been suggested that judicial sentencing be adapted to methods for monitoring and measuring brain activity, mostly in civil law for calculating moral damages, and in criminal law to individualize sentences.

In civil law, introducing new neuroscientific methods to "quantifying" alleged pain and damages would save time in procedural matters, solving and preventing legal disputes. Moreover, civil law applies a less rigorous burden of proof than criminal law: "the evidentiary rules will not apply in their full rigor, possibly making the admission of such evidence more likely." Procedural legal practice could thus be transformed more quickly in civil law than in criminal law.

In criminal law, the idea is to go from a retributivist conception of the law where criminals deserve their sentences, to a consequentialist conception of the law where considerations for consequences for the group, deterrence, prevention and treatment prevail. In this framework, supported by many scientists (e.g., Greene and Cohen, 2004; Sapolsky, 2004), the criminal is no longer a guilty person deserving sanctions but a sick individual to cure, and sometimes a simple danger for society to neutralize. Some pretend that sentences would then be more "human." However, Pardo and Patterson (2013), as well as Morse and Roskies, show that contrary to what we may believe, abandoning merits to justify sentences does not lead to softer sentences. On the contrary, "*(. . .) most of the most draconian aspects of punishment have been motivated by consequential concerns. Striking examples are recidivist sentencing enhancements, the approval of strict liability crimes, the "war on drugs". . . and mandatory minimum sentences. None of these can be retributively justified, and all punish disproportionately to desert*" (Morse and Roskies, 2013). The notion of deserved individual blame acts as a safeguard against the society's hegemonic temptation for security, in the name of which society is often prone, following a consequentialist approach, to sacrifice individual rights. Neuroscience, although certainly not sufficient to choose between legal conceptions, could nevertheless help us improve sentences in terms of efficiency by refining mental disorders or differential diagnosis. Again, neuroscience would not revolutionize law but improve already well-embedded practices (on the matter of potential future neurolaw revolutions, see Kolber, 2014).

Moreover, they give rise to the age-old ethical questions relating to the moral admissibility of certain physical treatments. Some scientists argue for attenuating immoral behaviors, such as racism and physical aggression, through TMS interventions or psychotropic drugs (Douglas, 2008). On the more consensual end of the scale, Coppola's propositions (Coppola, 2018) concern the use of predictive neuroscientific tools to evaluate recidivism rates[38], or the individualization of sentences to fit criminals neurobiology and facilitate social reinsertion.[39] However, the question that arises here, as it indeed has over the course of the history of criminal law, is to choose whether or not criminals should be corrected by means of physical interventions, or by education, punishment, etc. It has always been possible to cut a thief's arm, or to chemically castrate sexual delinquents. Sociology also presented itself as a good means to evaluate recidivism (Wootton, 1963). Neuroscience only counts here as another possibility on the long list of potential treatments for criminals [for a parallel drawn between the 1960s aversion therapies, as portrayed in *A Clockwork Orange* (Burge, Kubrick), and new techniques such as DBS and WBS, see McMillan, 2018]. Their admissibility leads the way to the procession of eternal ethical questions: the place for the accused's consent, physical integrity and identity, autonomy, retributivism and consequentialism, etc. (for a thorough discussion presenting both sides, for and against neurological interventions of criminals, see Birks and Douglas, 2018).

Enhanced Moral Agents

One original suggestion, instead of supporting a paradigm revolution or neuro-treatment, points toward "moral enhancement." The literature on this topic has arisen with the advent of new ways of enhancing one's cognitive capacities (may it be smart drugs, DBS etc.), and mostly deals with the main issue of the ethical permissibility of neurointerventions (see Persson and Savulescu, 2008, 2011, 2013; Harris, 2011; Douglas, 2013). Some authors delve specifically into the nexus between enhanced capacities and legal responsibility, questioning

[38]For example, the level of activity in the ACC might provide specific information as to whether an offender will be rearrested within 4 years of his release (Aharoni et al., 2013). Along the same line, a correlation has been found between reduced amygdala volume and increased risk for committing future violence in both young and adult males (e.g., Pardini et al., 2014; see Glenn and Raine, 2014; Coppola, 2018, for a review). Other important studies (Pustilnik, 2015, for a review) have characterized potential objective neural measures of how much subjective pain a subject is experiencing – which is important because the law's system of compensation in personal injury cases awards damages for pain based on mostly subjective assessment (Morse, 2019). These studies surely give a hint about the potential contributions that neuroscience may make to law in the future. Note that we do not insist on these potential contributions because the main focus of our paper is the foundation of criminal responsibility rather than the reliability of neuroscience-based methodology for e.g., predicting criminal behavior or better calculating damages.

[39]Neuroprediction might thus "*foster the implementation of alternative individualized sentences tending to offenders' actual social rehabilitation and social reintegration. Notably, neuroprediction could assist criminal justice systems to integrate current punitive policies and measures with socio-rehabilitative strategies, which could ultimately improve crime prevention and public safety without undermining the individual rights of offenders (. . .). An example of how neuroscience proves helpful in rehabilitative sentencing comes from Canada, where neurofeedback treatment programs have been tested on juvenile offenders*" (Coppola, 2018).

for example the duty to take enhancers in certain contexts (and the corollary liability for omissions), the breach of the standard of care that omitting to take enhancers could amount to, and the legal causal nexus between this type of omission and harm (see Goold and Maslen, 2015, for a discussion on those three points and a refutation that enhancers would give rise to these legal situations). Given the extensive literature on moral enhancement, we will only focus here on the influence of cognitive enhancement in determining criminal responsibility, and accordingly, on the validity of the underlying premise behind most claims relating to enhanced responsibility. That premise goes roughly as follows: if (criminal) responsibility is capacitarian and neuro-interventions can enhance our capacities, then those interventions could lead to an enhanced responsibility. In other words, responsibility would account for hypercapacity.

In a synthetic and systematic manner, Nicole Vincent explores the speculative question of responsibility enhancement, arguing specifically on the validity of the aforementioned premise (Vincent, 2013). She first exposes daily cases of responsibility assignments that follow greater capacities: when, for example, we say to a particularly mature child that disappoints us: "I expected more of you." She then answers 8 objections against the argument that responsibility accounts for hypercapacity, and demonstrates that enhanced capacities could lead to greater responsibility. Enhanced individuals could then be "*expected to satisfy higher standards*. . . *and they may even be deemed negligent or reckless for failure or refusal to do so, and possibly even sanctioned*" (Vincent, 2013, p. 329).

Intriguing though that idea may be, it is not immune from criticism. First, criminal responsibility, although capacitarian, is not *proportional* to an individual's capacities. Responsibility is attributed once certain criteria have been met: it is a threshold, not a scale[40]. The difference in severity across sentences is explained by the absence of certain criteria and not a partial fulfillment thereof. An act, in law, can be characterized as both "voluntary" and not intentional, but not as half voluntary and half intentional (such as manslaughter in Canada, corresponds to voluntary acts of violence without the intention of killing). The same goes for attenuating or aggravating circumstances: those only come into play once *mens rea* has been established. Against this objection, Vincent contends that although responsibility is a threshold, that threshold could be elevated through new cognitive enhancement techniques. Indeed, the "reasonable person" standard has evolved over time. There is a "reasonable person" for every place and time. To know what a future reasonable person will be for the western world is a sociological rather than legal question. In the hypothesis that the future reasonable person would have multiple brain implants, criminal law would remain unchallenged. Only the social norm would have changed. It is also worth noting that this new norm would only concern cases of objective responsibility (i.e., cases of omissions) and that actions would still be assessed through the lens of subjective responsibility (i.e., the subjective abilities to have a feeling of agency, to distinguish right from wrong, etc.). Finally, such an enhancement of the responsibility threshold, does not confirm, as Vincent seems to suggest, that "*responsibility tracks hypercapacity,*" but only that "*responsibility tracks capacity*" (which is a totally uncontroversial statement). Therefore, the so-called enhancement would not be considered enhanced at all, being the new standard.

Secondly, if we allow ourselves to speculate on a responsibility that would be proportional to capacities, we could only observe the disastrous and unfair consequences of such a notion. To be able to judge over-capable, or under-capable, individuals responsible for negligence, we need an objective standard to compare them to. There could not be the "reasonable person" single standard anymore, but a myriad of standards, the "more" or "less," "little" or "very" reasonable person. The multiplication of standards contradicts de facto the principle of equality in law, and would lead to segregated judicial orders for different classes of population. Besides, the matter of diagnosing hypercapacity remains delicate. Such a diagnosis could not be done at trial, since the accused would thereby never know the applicable standard until their first encounter with the law. Should we then test people every year during the whole of their lives in the eventuality that they be criminally charged? How would such tests work? Although this is somehow theoretically possible – for example, through behavioral modeling of developmental trajectories (e.g., Palminteri et al., 2016) –, it problematically ties such standards to the ungraspable, if not arbitrary, rhythm of scientific progress. Considering that the judges and the jury would have to always be on point concerning science's evolution, this policy seems impracticable.

To conclude, and extend beyond the initial scope of this section (i.e., focusing on the statement according to which criminal responsibility tracks hypercapacity), let us note that a version of enhanced responsibility already exists in our societies. Ministers, bosses, military superior officers are all people carrying a heavier responsibility tied to their functions. The weight of responsibility in those cases does not flow from greater capacities, but rather from the authority they exercise. Some of the literature on "moral enhancement" suggests that stronger individuals on a neurological level would be vested with some special authority and responsibility in their interactions with others. Neurological strength, however, gives you an authority that is primarily intimate, and not social: you exercise it on yourself, not on others. Should we then "biologise" the notion of authority in such a way that it extends to capacity? Would we not thereby void its meaning as a social influence that we accept at the expense of a greater vulnerability to society's demands? Is it not fairer for a social institution, by which people judge each other, to lie in the choices individuals make in relation to one another? To her credit, Vincent recognizes the importance of choices (that she addresses through the angle of *consent* to responsibility). She nonetheless considers them as one out of many aspects of responsibility. In this regard, we disagree: criminal responsibility, at least in liberal democracies, is rooted in a social contract.

[40]This threshold is relatively simple to reach and does not require extraordinary morality (an ability to distinguish between right and wrong, to perceive the world correctly, to act according to one's own intentions, etc.).

Individual choices are not simple considerations, but the very foundations (and/or justification) of it all.

The question that remains, and indeed is a constant concern, for theorists of enhanced responsibility is this: should foundations of responsibility be "neurologised"? It seems obvious that we regard judging each other's actions as something that is beneficial to society, but what of judging each other's biological make up?

CONCLUSION

In the course of our analysis, we have defined criminal responsibility as an essentially practical concept independent from free will and other metaphysical questions. Hence, criminal responsibility is immune from debates on determinisms and their affiliated answers. We have recalled that the current and retributivist model of criminal responsibility affords a central place to the individual in relation with the sentence. While asking for an individual's reasons to act, it treats that individual as a person who deserves blame, but also dignity. Questioning a person's reasons to act and feeling of responsibility also serves the purpose of evaluating the fairness of the conditions given by society for making a choice. That model is anchored in current popular beliefs regarding accountability and the promotion of certain values. If traditional neuroscience disciplines want to revolutionize law, they cannot simply establish facts. On their own, without any ideological aim, they cannot substantially modify normative practices. They must also engage in a political justification of the changes being called for, convince populations, and take into consideration the ensuing consequences. In turn, this approach must acknowledge and deal with technical, interpretative and legal obstacles that limit the uniform application of neuroscience. Far from a revolution, neuroscience proves to be more beneficial when entering in a subtle dialogue with the law in order to assist the truth-seeking function of the courts. In other words, neuroscience's greatest potential with respect to the law lies less in assessing the degree of responsibility of an accused than in reconstructing a state of affairs and determining what the implications of that state of affairs may be with respect to the accuracy of allegations.

While neurolaw often evokes the neuroscientification of law, it could more properly refer to the juridification of neuroscience, i.e., legal thinking that would integrate and apply scientific discoveries to criminal justice.

AUTHOR CONTRIBUTIONS

AB and VC wrote the manuscript with equal contributions.

FUNDING

This work was supported by the Agence Nationale de la Recherche (ANR) grants ANR-17-EURE-0017 (Frontiers in Cognition), ANR-10-IDEX-0001-02 PSL* (program "Investissements d'Avenir"), and ANR-16-CE37-0012-01.

ACKNOWLEDGMENTS

We thank Melissa Gregg, Patrick Haggard, Stefan Hnatiuk, and Nura Sidarus for comments and useful discussions regarding earlier versions of this manuscript.

REFERENCES

Aharoni, E., Vincent, G. M., Harenski, C. L., Calhoun, V. D., Sinnott-Armstrong, W., Gazzaniga, M. S., et al. (2013). Neuroprediction of future rearrest. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6223–6228. doi: 10.1073/pnas.1219302110

Aizawa, K. (2009). Neuroscience and multiple realization: a reply to bechtel and mundale. *Synthese* 167, 493–510. doi: 10.1007/s11229-008-9388-5

Anderson, M. L. (2010). Neural reuse: a fundamental organizational principle of the brain. *Behav. Brain Sci.* 33, 245–266. doi: 10.1017/S0140525X10000853

Andics, A., Gábor, A., Gácsi, M., Faragó, T., Szabó, D., and Miklósi, Á. (2016). Neural mechanisms for lexical processing in dogs. *Science* 353, 1030–1032. doi: 10.1126/science.aaf3777

Andics, A., Gábor, A., Gácsi, M., Faragó, T., Szabó, D., and Miklósi, Á. (2017). Erratum for the report "Neural mechanisms for lexical processing in dogs" by A. Andics, A. Gábor, M. Gácsi, T. Faragó, D. Szabó, Á. Miklósi. *Science* 356:eaan3276. doi: 10.1126/science.aan3276

Baertschi, B. (2009). *La neuroéthique: ce que les neurosciences font à nos conceptions morales*. Paris: Editions La Découverte.

Barclay, P., and Raihani, N. (2016). Partner choice versus punishment in human prisoner's dilemmas. *Evol. Hum. Behav.* 37, 263–271. doi: 10.1016/j.evolhumbehav.2015.12.004

Baumard, N., and Sperber, D. (2010). Weird people, yes, but also weird experiments. *Behav. Brain Sci.* 33, 84–85. doi: 10.1017/S0140525X10000038

Birks, D., and Douglas, T. (2018). *Treatment for Crime. Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford: Oxford University Press.

Bouton, S., Chambon, V., Tyrand, R., Guggisberg, A. G., Seeck, M., Karkar, S., et al. (2018). Focal versus distributed temporal cortex activity for speech sound category assignment. *Proc. Natl. Acad. Sci. U.S.A.* 115, E1299–E1308. doi: 10.1073/pnas.1714279115

Brandom, R. (2015). *Is « Brain Fingerprinting » a Breakthrough or a Sham? The Verge*. Available at: http://www.theverge.com/2015/2/2/7951549/brain-fingerprinting-technology-unproven-courtroom-science-farwell-p300 (accessed February 2, 2015).

Brass, M., and Haggard, P. (2008). The what, when, whether model of intentional action. *Neuroscientist* 14, 319–325. doi: 10.1177/1073858408317417

Buchy, L., Woodward, T. S., and Liotti, M. (2007). A cognitive bias against disconfirmatory evidence (BADE) is associated with schizotypy. *Schizophr. Res.* 90, 334–337. doi: 10.1016/j.schres.2006.11.012

Capestany, B. H., and Harris, L. T. (2014). Disgust and biological descriptions bias logical reasoning during legal decision-making. *Soc. Neurosci.* 9, 265–277. doi: 10.1080/17470919.2014.892531

Chambon, V., Domenech, P., Pacherie, E., Koechlin, E., Baraduc, P., and Farrer, C. (2011). What are they up to? The role of sensory evidence and prior knowledge in action understanding. *PLoS One* 6:e17133. doi: 10.1371/journal.pone.0017133

Chambon, V., Filevich, E., and Haggard, P. (2014a). "What is the human sense of agency, and is it Metacognitive?," in *The Cognitive Neuroscience of Metacognition*, eds S. M. Fleming and C. D. Frith (Heidelberg: Springer), 321–342. doi: 10.1007/978-3-642-45190-4_14

Chambon, V., Sidarus, N., and Haggard, P. (2014b). From action intentions to action effects: how does the sense of agency come about? *Front. Hum. Neurosci.* 8:320. doi: 10.3389/fnhum.2014.00320

Chisholm, R. (1976). "The Agent as Cause," in *Action Theory*, eds M. Brand and D. Walton (Dordrecht: D.Reidel), 199–211. doi: 10.1007/978-94-010-9074-2_12

Claar v. Burlington (1994). Claar v. Burlington N.R.R., 29 F.3d 499 (9th Cir. 1994). doi: 10.1007/978-94-010-9074-2_12

Coppola, F. (2018). Mapping the brain to predict antisocial behaviour: new frontiers in neurocriminology,'new'challenges for criminal justice. *UCL J. Law Jurisprud. Spec. Issue* 1, 103–126.

Cullen v. Pinholster (2011). *Cullen v. Pinholster 590 F. 3d 651 (2011)*.

Cunnington, R., Windischberger, C., Deecke, L., and Moser, E. (2003). The preparation and readiness for voluntary movement: a high-field event-related fMRI study of the Bereitschafts-BOLD response. *Neuroimage* 20, 404–412. doi: 10.1016/s1053-8119(03)00291-x

Dahan-Katz, L. (2013). "The implications of heuristics and biases research on moral and legal responsibility," in *Neuroscience and Legal Responsibility,* ed. N. A. Vincent (New York: Oxford University Press), 135–161.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proc. Natl. Acad. Sci. U.S.A.* 108, 6889–6892. doi: 10.1073/pnas.1018033108

Daubert v. Merrell Dow Pharmaceuticals, Inc (1995). *Daubert v. Merrell Dow Pharmaceuticals, Inc., 43 F 1311 (9th Cir. 1995)*.

Davis, T., and Poldrack, R. A. (2013). Measuring neural representations with fMRI: practices and pitfalls. *Ann. N. Y. Acad. Sci.* 1296, 108–134. doi: 10.1111/nyas.12156

Dawkins, R. (2006). *Let's all stop beating Basil's car", The Edge Annual Question 2006, 'What is your Dangerous Idea?* Available at: www.edge.org/q2006/q06_9.html#dawkins

Douglas, T. (2008). Moral enhancement. *J. Appl. Philos.* 25, 228–245.

Douglas, T. (2013). Moral enhancement via direct emotion modulation: a reply to John Harris. *Bioethics* 27, 160–168. doi: 10.1111/j.1467-8519.2011.01919.x

Drew, P. J., Winder, A. T., and Zhang, Q. (2018). Twitches, blinks, and fidgets: important generators of ongoing neural activity. *Neuroscientist* [Epub ahead of print].

Eitam, B., and Haggard, P. (2015). *The Sense of Agency*. Oxford: Oxford University Press.

Eklund, A., Knutsson, H., and Nichols, T. E. (2018). Cluster failure revisited: impact of first level design and data quality on cluster false positive rates. *Hum. Brain Mapp.* arXiv:1804.03185.

Falk, E. B., Hyde, L. W., Mitchell, C., Faul, J., Gonzalez, R., Heitzeg, M. M., et al. (2013). What is a representative brain? Neuroscience meets population science. *Proc. Natl. Acad. Sci. U.S.A.* 110, 17615–17622. doi: 10.1073/pnas.1310134110

Farahany, N. A. (2012). Incriminating thoughts. *SLR.* 64, 351–408.

Farwell, L. A. (2012). Brain fingerprinting: a comprehensive tutorial review of detection of concealed information with event-related brain potentials. *Cogn. Neurodyn.* 6, 115–154. doi: 10.1007/s11571-012-9192-2

Farwell, L. A., and Donchin, E. (1991). The truth will out: interrogative polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology* 28, 531–547. doi: 10.1111/j.1469-8986.1991.tb01990.x

Farwell, L. A., and Smith, S. S. (2001). Using brain MERMER testing to detect knowledge despite efforts to conceal. *J. Forensic Sci.* 46, 135–143.

Faulkner, P. (2007). What is wrong with lying? *Philos. Phenomenol. Res.* 75, 535–557. doi: 10.1111/j.1933-1592.2007.00092.x

Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. *Nature* 415, 137–140. doi: 10.1038/415137a

Frankfurt, H. G. (1988). "Freedom of the Will and the Concept of a Person," in *What is a person? Contemporary Issues in Biomedicine, Ethics, and Society*, ed. M. F. Goodman (Clifton, NJ: Humana Press), 127–144. doi: 10.1007/978-1-4612-3950-5_6

Frisch, M. (2014). *Causal Reasoning in Physics*. Cambridge: Cambridge University Press.

Frith, C. D., and Haggard, P. (2018). Volition and the brain–revisiting a classic experimental study. *Trends Neurosci.* 41, 405–407. doi: 10.1016/j.tins.2018.04.009

Frye v. United States (1923). *F 1013 (D.C.Circ. 1923)*.

General Elec, Co. v. Joiner (1997). *General Elec. Co. v. Joiner, 522 U.S. 136, 146 (U.S. 1997)*.

George, M., Kozel, F., and Bohning, D. (2006). Functional magnetic resonance imaging guided transcranial magnetic stimulation deception inhibitor. U.S. Patent Application No 10/521,373

Gigerenzer, G., and Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychol. Rev.* 103, 650–669. doi: 10.1037/0033-295x.103.4.650

Gigerenzer, G., and Selten, R. (2002). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT press.

Glenn, A. L., and Raine, A. (2014). Neurocriminology: implications for the punishment, prediction and prevention of criminal behaviour. *Nat. Rev. Neurosci.* 15, 54–63. doi: 10.1038/nrn3640

Goold, I., and Maslen, H. (2015). "Responsibility Enhancement and the Law of Negligence," in *Handbook of Neuroethics*, eds J. Clausen and N. Levy (Dordrecht: Springer).

Greene, J., and Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 359, 1775–1785.

Guala, F. (2012). Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* 35, 1–15. doi: 10.1017/S0140525X11000069

Haggard, P. (2017). Sense of agency in the human brain. *Nat. Rev. Neurosci.* 18, 196–207. doi: 10.1038/nrn.2017.14

Haggard, P., and Chambon, V. (2012). Sense of agency. *Curr. Biol.* 22, R390–R392.

Han, H., and Glenn, A. L. (2018). Evaluating methods of correcting for multiple comparisons implemented in SPM12 in social neuroscience fMRI studies: an example from moral psychology. *Soc. Neurosci.* 13, 257–267. doi: 10.1080/17470919.2017.1324521

Harrington v. State (2003). *Harrington v. State, 659 N.W.2d 509 (Iowa 2003)*.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics* 25, 102–111. doi: 10.1111/j.1467-8519.2010.01854.x

Hart, H. L. A. (1968). *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Oxford University Press.

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., et al. (2009). Adaptive rationality: an evolutionary perspective on cognitive bias. *Soc. Cogn.* 27, 733–763. doi: 10.1521/soco.2009.27.5.733

Haselton, M. G., and Nettle, D. (2006). The paranoid optimist: an integrative evolutionary model of cognitive biases. *Pers. Soc. Psychol. Rev.* 10, 47–66. doi: 10.1207/s15327957pspr1001_3

Haushalter, J. L. (2018). Neuronal testimonial: brain-computer interfaces and the law. *Vand. Law Rev.* 71:1365.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Beyond WEIRD: towards a broad-based behavioral science. *Behav. Brain Sci.* 33, 111–135. doi: 10.1017/S0140525X10000725

Hu, C. P., Jiang, X., Jeffrey, R., and Zuo, X. N. (2018). Open science as a better gatekeeper for science and society: a perspective from neurolaw. *Sci. Bull.* 63, 1529–1531. doi: 10.1016/j.scib.2018.11.015

Jeannerod, M. (2001). Neural simulation of action: a unifying mechanism for motor cognition. *Neuroimage* 14, S103–S109.

Kaufmann, P. M. (2013). Neuropsychologist experts and neurolaw: cases, controversies, and admissibility challenges. *Behav. Sci. Law* 31, 739–755. doi: 10.1002/bsl.2085

Kedia, G., Harris, L., Lelieveld, G.-J., and van Dillen, L. (2017). From the brain to the field: the applications of social neuroscience to economics, health and law. *Brain Sci.* 7:94. doi: 10.3390/brainsci7080094

Keech v. Commonwealth (1989). *Keech v. Commonwealth, 9 Va. App.,386 S.E.2d 1989*.

Klaming, L., and Haselager, P. (2013). Did my brain implant make me do it? Questions raised by DBS regarding psychological continuity, responsibility for action and mental competence. *Neuroethics* 6, 527–539. doi: 10.1007/s12152-010-9093-1

Kolber, A. J. (2014). Will there be a neurolaw revolution. *Indiana Law J.* 89:807.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93, 480–490. doi: 10.1016/j.neuron.2016.12.041

Kulakova, E., Khalighinejad, N., and Haggard, P. (2017). I could have done otherwise: availability of counterfactual comparisons informs the sense of agency. *Conscious. Cogn.* 49, 237–244. doi: 10.1016/j.concog.2017.01.013

Kumho Tire Co. v. Carmichael (1999). *Kumho Tire Co. v. Carmichael, 119 U.S. 1167, 1175 (U.S. 1999)*.

Lakens, D. (2017). *Impossibly Hungry Judges. The 20% Statistician*. Available at: https://daniellakens.blogspot.com/2017/07/impossibly-hungry-judges.html (accessed July 3, 2017).

Levy, N. (2014). Is neurolaw conceptually confused? *J. Ethics* 18, 171–185. doi: 10.1007/s10892-014-9168-z

Libet, B. W. (1999). Do we have free will? *J. Conscious. Stud.* 6, 47–57.

Ligneul, R., Sescousse, G., Barbalat, G., Domenech, P., and Dreher, J.-C. (2013). Shifted risk preferences in pathological gambling. *Psychol. Med.* 43, 1059–1068. doi: 10.1017/S0033291712001900

Loftus, E. F., and Palmer, J. C. (1974). Reconstruction of automobile destruction: an example of the interaction between language and memory. *J. Verbal Learn. Verbal Behav.* 13, 585–589. doi: 10.1080/17470218.2016.1237980

Lukács, G., Weiss, B., Dalos, V. D., Kilencz, T., Tudja, S., and Csifcsák, G. (2016). The first independent study on the complex trial protocol version of the P300-based concealed information test: corroboration of previous findings and highlights on vulnerabilities. *Int. J. Psychophysiol.* 110, 56–65. doi: 10.1016/j.ijpsycho.2016.10.010

Maibom, H. L. (2008). The mad, the bad, and the psychopath. *Neuroethics* 1, 167–184. doi: 10.1007/s12152-008-9013-9

Matravers, M., and Cocoru, A. (2014). "Revisiting the Hart/Wootton Debate on Responsibility," in *Hart on Responsibility*, ed. C. G. Pulman (New York, NY: Palgrave Macmillan).

McMillan, J. (2018). "Containing Violence and Controlling Desire," in *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*, eds D. Birks and T. Douglas (Oxford: Oxford University Press).

Meijer, E. H., Ben-Shakhar, G., Verschuere, B., and Donchin, E. (2013). A comment on farwell (2012): brain fingerprinting: a comprehensive tutorial review of detection of concealed information with event-related brain potentials. *Cogn. Neurodyn.* 7, 155–158. doi: 10.1007/s11571-012-9217-x

Meijer, E. H., Selle, N. K., Elber, L., and Ben-Shakhar, G. (2014). Memory detection with the concealed information test: a meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology* 51, 879–904. doi: 10.1111/psyp.12239

Mele, A. R. (2009). *Effective Intentions: The Power of Conscious Will*. Oxford: Oxford University Press.

Meynen, G. (2014). Neurolaw: neuroscience, ethics, and law. review essay. *Ethical Theory Moral Pract.* 17, 819–829. doi: 10.1007/s10677-014-9501-4

Morse, S. J. (2002). Uncontrollable urges and irrational people. *Va. Law Rev.* 88, 1025–1078.

Morse, S. J. (2006). *Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note (SSRN Scholarly Paper No. ID 896753)*. Rochester, NY: Social Science Research Network.

Morse, S. J. (2019). "Neurohype and the law: A cautionary tale," in *Casting Light on the Dark Side of Brain Imaging*, eds A. Raz and R. T. Thibault (London: Academic Press), 31–35.

Morse, S. J., and Roskies, A. L. (eds) (2013). *A Primer on Criminal Law and Neuroscience: A Contribution of the Law and Neuroscience Project, Supported by the MacArthur Foundation*. New York, NY: Oxford University Press.

Nahm, M., Rousseau, D., and Greyson, B. (2017). Discrepancy between cerebral structure and cognitive functioning: a review. *J. Nerv. Ment. Dis.* 205, 967–972. doi: 10.1097/NMD.0000000000000752

Nahmias, E., Shepard, J., and Reuter, S. (2014). It's OK if 'my brain made me do it': people's intuitions about free will and neuroscientific prediction. *Cognition* 133, 502–516. doi: 10.1016/j.cognition.2014.07.009

Nahmias, E., and Thompson, M. (2014). "A naturalistic vision of free will," in *Current Controversies in Experimental Philosophy*, eds E. O'Neill and E. Machery (London: Routledge), 86–103.

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175–220. doi: 10.1037//1089-2680.2.2.175

Pacherie, E. (2008). The phenomenology of action: a conceptual framework. *Cognition* 107, 179–217. doi: 10.1016/j.cognition.2007.09.003

Pallarés-Dominguez, D., and Esteban, E. G. (2016). The ethical implications of considering neurolaw as a new power. *Ethics Behav.* 26, 252–266. doi: 10.1080/10508422.2015.1012763

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., et al. (2002). The role of individual differences in the accuracy of confidence judgments. *J. Gen. Psychol.* 129, 257–299.

Palminteri, S., Kilford, E. J., Coricelli, G., and Blakemore, S.-J. (2016). The computational development of reinforcement learning during adolescence. *PLoS Comput. Biol.* 12:e1004953. doi: 10.1371/journal.pcbi.1004953

Pardini, D. A., Raine, A., Erickson, K., and Loeber, R. (2014). Lower amygdala volume in men is associated with childhood aggression, early psychopathic traits, and future violence. *Biol. Psychiatry* 75, 73–80. doi: 10.1016/j.biopsych.2013.04.003

Pardo, M. S. (2018). *Lying, Deception, and fMRI: A Critical Update. Neurolaw and Responsibility for Action*. Cambridge, MA: Cambridge University Press.

Pardo, M. S., and Patterson, D. (2013). *Minds, Brains, and Law: The Conceptual Foundations of Law and Neuroscience*. Oxford: Oxford University Press.

Parent, H. (2008). *Traité de droit criminel, Tome I - L'imputabilité*, 3ème Edn, Vol. 1. Montréal: Thémis.

Pedersen, E. J., McAuliffe, W. H., and McCullough, M. E. (2018). The unresponsive avenger: more evidence that disinterested third parties do not punish altruistically. *J. Exp. Psychol. Gen.* 147, 514–544. doi: 10.1037/xge0000410

Persson, I., and Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *J. Appl. Philos.* 25, 162–177. doi: 10.1111/j.1468-5930.2008.00410.x

Persson, I., and Savulescu, J. (2011). "Unfit for the future? Human nature, scientific progress, and the need for moral enhancement," in *Enhancing Human Capacities*, eds J. Savulescu, R. ter Meulen, and G. Kahane (Oxford: Wiley-Blackwell).

Persson, I., and Savulescu, J. (2013). Getting moral enhancement right: the desirability of moral bioenhancement. *Bioethics* 27, 124–131. doi: 10.1111/j.1467-8519.2011.01907.x

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron* 72, 692–697. doi: 10.1016/j.neuron.2011.11.001

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018

Pustilnik, A. C. (2015). Imaging brains, changing minds: how pain neuroimaging can inform the law. *Ala. L. Rev.* 66:1099.

R. v. Byrne (1960). *2 QB 396*.

R c. Gibson (2001). *R c. Gibson, 2001 153 C.C.C. (3d) 465*.

Reimer, M. (2008). Psychopathy without (the Language of) disorder. *Neuroethics* 1, 185–198. doi: 10.1007/s12152-008-9017-5

Ritchie, J. B., and Carlson, T. A. (2016). Neural decoding and "inner" psychophysics: a distance-to-bound approach for linking mind, brain, and behavior. *Front. Neurosci.* 10:190. doi: 10.3389/fnins.2016.00190

Ritchie, J. B., Kaplan, D. M., and Klein, C. (2017). Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *Br. J. Philos. Sci.* 70, 581–607. doi: 10.1093/bjps/axx023

Roper v. Simmons (2005). *Roper v. Simmons, 543 U.S. 551 (2005)*.

Rosen, J. (2007). *The Brain on the Stand. The New York Times*. Available at: http://www.nytimes.com/2007/03/11/magazine/11Neurolaw.t.html (accessed March 11, 2017).

Rosenfeld, J. P. (2005). Brain fingerprinting: a critical analysis. *Sci. Rev. Ment. Health Pract.* 4, 20–37.

Roskies, A. (2004). *A Case Study of Neuroethics: the Nature of Moral Judgment*. Oxford: Oxford University Press.

Russell, B. (1912). On the notion of cause. *Proc. Aristotelian Soc.* 13, 1–26.

Sapolsky, R. M. (2004). The frontal cortex and the criminal justice system. *Philos. Trans. R. Soc. B Biol. Sci.* 359, 1787–1796.

Satel, S., and Lilienfeld, S. O. (2013). *Brainwashed: The Seductive Appeal of Mindless Neuroscience*. New York, NY: Basic Civitas Books.

Schimmack, U., Heene, M., and Kesavan, K. (2017). *Reconstruction of a Train Wreck: How Priming Research Went off the Rails. Replicability Index*. Available at: https://replicationindex.wordpress.com/2017/02/02/reconstruction-of-a-train-wreck-how-priming-research-went-of-the-rails/#comment-1454 (accessed February 2, 2017).

Schmeiser, B., Zentner, J., Steinhoff, B. J., Schulze-Bonhage, A., Kogias, E., Wendling, A.-S., et al. (2017). Functional hemispherectomy is safe and effective in adult patients with epilepsy. *Epilepsy Behav.* 77, 19–25. doi: 10.1016/j.yebeh.2017.09.021

Schurger, A., Sitt, J. D., and Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proc. Natl. Acad. Sci. U.S.A.* 109, E2904–E2913.

Searle, J. R. (1984). *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

Seymour, K., Clifford, C. W., Logothetis, N. K., and Bartels, A. (2009). The coding of color, motion, and their conjunction in the human visual cortex. *Curr. Biol.* 19, 177–183. doi: 10.1016/j.cub.2008.12.050

Shaver, K. G. (2012). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. Berlin: Springer Science & Business Media.

Sheehan v. Daily Racing Form, Inc. (1997). *(104)*F.3d 940, 942 (7th Cir. 1997).

Shepard, J., and O'Grady, A. (2017). What kinds of alternative possibilities are required of the folk concept of choice? *Conscious. Cogn.* 48, 138–148. doi: 10.1016/j.concog.2016.11.005

Shepperd, J., Malone, W., and Sweeny, K. (2008). Exploring causes of the self-serving bias. *Soc. Pers. Psychol. Compass* 2, 895–908. doi: 10.1111/j.1751-9004.2008.00078.x

Shepard, J., and Reuter, S. (2012). Neuroscience, choice, and the free will debate. *AJOB Neurosci.* 3, 7–11. doi: 10.1080/21507740.2012.694390

Sinnott-Armstrong, W., and Nadel, L. (eds) (2010). *Conscious Will and Responsibility: A Tribute to Benjamin Libet*. New York, NY: Oxford University Press.

Spranger, T. (ed.) (2012). *International Neurolaw: A Comparative Analysis*. Heidelberg: Springer-Verlag.

Synofzik, M., Vosgerau, G., and Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Conscious. Cogn.* 17, 219–239. doi: 10.1016/j.concog.2007.03.010

Van Dijk, K. R., Sabuncu, M. R., and Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage* 59, 431–438. doi: 10.1016/j.neuroimage.2011.07.044

Van Horne, W. A. (1981). Prolegomenon to a theory of deception. *Philos. Phenomenol. Res.* 42, 171–182.

Vincent, N. (2013). "Enhancing responsibility," in *Neuroscience and Legal Responsibility*, ed. N. Vincent (New York, NY: Oxford University Press).

Vincent, N. A. (2010). On the relevance of neuroscience to criminal responsibility. *Crim. Law Philos.* 4, 77–98. doi: 10.1007/s11572-009-9087-4

Vincent, N. A. (2011). Neuroimaging and responsibility assessments. *Neuroethics* 4, 35–49. doi: 10.1007/s12152-008-9030-8

Waller, R. (2012). Beyond button presses: the neuroscience of free and morally appraisable actions. *Monist* 95, 441–462. doi: 10.5840/monist201295323

Wang, H. X., Merriam, E. P., Freeman, J., and Heeger, D. J. (2014). Motion direction biases and decoding in human visual cortex. *J. Neurosci.* 34, 12601–12615. doi: 10.1523/JNEUROSCI.1034-14.2014

Wegner, D. M. (2002). *The illusion of Conscious Will*. Cambridge, MA: MIT Press.

Weinshall-Margel, K., and Shapard, J. (2011). Overlooked factors in the analysis of parole decisions. *Proc. Natl. Acad. Sci. U.S.A.* 108, E833–E833.

Wittenberg, G. F. (2010). Experience, cortical remapping, and recovery in brain disease. *Neurobiol. Dis.* 37, 252–258. doi: 10.1016/j.nbd.2009.09.007

Woo, C.-W., Krishnan, A., and Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage* 91, 412–419. doi: 10.1016/j.neuroimage.2013.12.058

Wootton, B. (1963). *Crime and the Criminal Law: Reflections of a Magistrate and Social Scientist*. London: Stevens.

Yaffe, G. (2013). *Are Addicts Akratic? Interpreting the Neuroscience of Reward*. Oxford: Oxford University Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

REVIEW
published: 13 March 2020
doi: 10.3389/fpsyg.2020.00390

Re-wiring Guilt: How Advancing Neuroscience Encourages Strategic Interventions Over Retributive Justice

Nathaniel E. Anderson[1]* *and Kent A. Kiehl*[1,2]

[1] *The Mind Research Network, Albuquerque, NM, United States,* [2] *Departments of Psychology, Neuroscience, and Law, University of New Mexico, Albuquerque, NM, United States*

The increasing visibility of neuroscience employed in legal contexts has rightfully prompted critical discourse regarding the boundaries of its utility. High profile debates include some extreme positions that either undermine the relevance of neuroscience or overstate its role in determining legal responsibility. Here we adopt a conciliatory attitude, reaffirming the current value of neuroscience in jurisprudence and addressing its role in shifting normative attitudes about culpability. Adopting a balanced perspective about the interaction between two dynamic fields (science and law) allows for more fruitful consideration of practical changes likely to improve the way we engage in legal decision-making. Neuroscience provides a useful platform for addressing nuanced and multifaceted deterministic factors promoting antisocial behavior. Ultimately, we suggest that shifting normative attitudes about culpability vis-à-vis advancing neuroscience are not likely to promote major changes in the way we assign legal responsibility. Rather, it helps us to shed our harshest retributivist instincts in favor of more pragmatic strategies for combating the most conspicuous patterns promoting mass incarceration and recidivism.

Keywords: neurolaw, neuroscience, jurisprudence, free will, determinism, culpability, intervention

OPEN ACCESS

Edited by:
José M. Muñoz, Universidad Europea de Valencia, Spain

Reviewed by:
Federico Gustavo Pizzetti, University of Milan, Italy
Mirko Daniel Garasic, Libera Università Maria SS. Assunta, Italy

***Correspondence:**
Nathaniel E. Anderson
nanderson@mrn.org

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *11 December 2019*
Accepted: *20 February 2020*
Published: *13 March 2020*

Citation:
Anderson NE and Kiehl KA (2020) Re-wiring Guilt: How Advancing Neuroscience Encourages Strategic Interventions Over Retributive Justice. Front. Psychol. 11:390. doi: 10.3389/fpsyg.2020.00390

INTRODUCTION

Increasing attention is being devoted to the emerging roles of neuroscience in legal decision-making, both in academic settings and in the courtroom. Among these roles is the growing influence neuroscience has in reinforcing more deterministic models of human decision-making and behavior. In a deliberate attempt to (over)simplify this complex landscape, it seems that there are roughly two camps in these conversations. The first includes those who promote the idea that neuroscience has mostly "disproved" the existence of free will, which subverts some of our ordinary notions of accountability. Prominent voices in popular media have indeed heralded the *end of free will* (Harris, 2012; Cave, 2016), calling into question our most basic presumptions about the legitimacy of punitive justice (Burns and Bechara, 2007; Sapolsky, 2017). As a consequence, the criminal justice system, which punishes people on a now-baseless presumption of freedom and agency, has been foundationally undermined and must therefore be replaced immediately with something more enlightened and fair. Expectedly, such claims have incited substantial opposition and motivated counter-arguments aimed at substantiating traditional views of legal responsibility

and the enduring value of retributive sanctions against criminal actions. This opposing camp includes those who claim that free will (whether it exists or not) is largely irrelevant to basic notions of legal responsibility, and neuroscience has little to no relevance for assessing guilt or any ordinary sense of civic accountability. As such, the *status quo* can be safely perpetuated, and the law, as it stands, remains unfettered by trifling nuisances of pre-determined actions. Perhaps as a kind of contrecoup effect, these counter arguments often involve rebuffing the very relevance of neuroscience in the legal process more generally (Morse, 2006; Pardo and Patterson, 2010; Chambon and Bigenwald, 2019). Of course these descriptions are composite caricatures of many subtler perspectives available in this growing academic conversation, e.g. Vincent (2013). Still, many of the arguments we read these days overlap substantially and obviously with one of these two extreme positions. Without depreciating the zeitgeist of this revolution or its opponents, we recognize the need for some conservatism in advancing pragmatic attitudes about how the legal system might change in the wake of advancing neuroscience. We therefore set out here to willfully explore the vast middle ground between seemingly extreme perspectives in this conversation. We ultimately promote three main theses related to *Neurolaw* and its inevitable progress.

Neuroscience Has Firmly Established Its Place in Jurisprudence

Neuroscience already plays a prominent role in legal proceedings. This trend seems likely to increase rather than decrease; however, this development should not be alarming. As judges and juries are faced with the challenge of weighing this evidence in their decision-making, we have a responsibility to make this process as transparent as possible. This involves promoting better science and better education to judges and legal counsel about these data's interpretation, limitations, and best practices in quality assurance and analytic strategies.

Our Normative Understanding of Free Will and Culpability Is Changing

Our understanding of human behavior and free will has steadily incorporated more contributions from science throughout history. The role of neuroscience only represents a recent and specific extension of this progress. Increasingly deterministic models of behavior need not cripple our aim to hold people responsible for their actions, but they arguably drive changes in our normative view of culpability and what constitutes justice.

Our Legal System Is Evolving, Not Static

Recent changes in our legal system highlight evolving standards in normative judgments regarding the relative value of retributive and rehabilitative interventions. Neuroscience provides a platform to re-assess the value of primarily punitive systems that have historically done little to remedy mental health and social issues that perpetuate high incarceration rates. Rather than eroding jurisprudence, this has the potential to inform more effective policies that serve our society in progressive ways.

NEUROSCIENCE HAS FIRMLY ESTABLISHED ITS PLACE IN JURISPRUDENCE

Critical evaluations of the role of neuroscience (and particularly neuroimaging) in legal proceedings abound in both academic publications and the popular media (Brown and Murphy, 2010; Eagleman, 2011a; Morse, 2015; Gonzalez, 2017). While the tone of these pieces can range from cautionary to insolent, they attend to a common issue of what is frequently described as a *meteoric rise* in the consideration of neuroscience-based evidence in courtroom decision-making. They frequently highlight perceived negative consequences of this trend, and some suggest the limited relevance of neuroimaging in court overall. Here we argue that neuroscience and neuroimaging in particular have already established their place in legal proceedings, which is unlikely to subside. A more practical approach for exercising caution in its application will be to improve stakeholders' understanding of the strengths and limitations of these techniques which includes educating lawyers, judges, and the general public. Educated adoption of neuroscience in legal settings is a practical and realistic solution to any perceived hazards it engenders.

The characterization of a *meteoric rise* of neuroscience used in court carries with it a somewhat menacing connotation that may not be wholly justified. While estimates have suggested an approximate doubling of cases that consider neuroscience data as legal evidence in the past decade (Catley and Claydon, 2016; Farahany, 2016), this is not out of step with its rise in clinical and research settings over the same period (Yeung et al., 2017). Further, in contrast to the tone of many commentaries, this steady increase has not occurred unexpectedly, overwhelming courts with claims that its practitioners cannot fairly evaluate. One of the first considerations of brain imaging as evidence in court occurred over 35 years ago in the high profile trial of John Hinckley Jr. for the attempted assassination of President Ronald Reagan (United States vs. Hinckley, 1982[1]). Brain scans were used in conjunction with other clinical evidence to support Hinckley's diagnosis of schizophrenia. The brain imaging was not foundational to his diagnosis, but served the purpose of grounding his claims of mental illness in a physical domain (as opposed to purely "psychological") – an important educational element for jurors in the early 1980s. This context remains among the most influential roles of neuroscience in contemporary jurisprudence, where judges and juries must inevitably weigh the "legitimacy" of health claims and related assertions that remain difficult to account for objectively (e.g., chronic pain, psychiatric disorders). Hinckley was found not guilty by reason of insanity and was committed to a secure hospital for the mentally ill.

Since then, the application of neuroscience evidence in court has increased, but not on a scale that is out of step with advancing scientific knowledge and improved practical utility. Any notions that the rise in neuroscience has happened too quickly for courts to implement its data sagaciously are likely misguided. Several

[1] *United States v. Hinckley*, 525 F. Supp. 1342 (D.C. 1981).

reports have objectively evaluated this changing landscape. Using broad inclusive criteria, it has been estimated that neuroscience evidence is considered in less than 1% of all criminal proceedings (based on court of appeals data) and only about 5% of murder trials dating back to 2005 (Catley and Claydon, 2016; De Kogel and Westgeest, 2016; Farahany, 2016). Contexts that have seen more pronounced increases include evaluation of competency to stand trial, capital cases, and appeals for mitigation of punishment. An observation made frequently across studies is that the rise in neuroscience evidence appears most striking in high-stakes cases. While these numbers are steadily increasing, they do so in-step with advancing scientific understanding and improved potential to inform judges and juries about a number of relevant aspects of mental health, within standard limits of due process – and this context is important. As with any emerging technology, its relevance in legal proceedings must be evaluated carefully with established evidentiary standards (Gaudet, 2011) as courts learn to integrate and adapt to progress in clinical neuroscience. Nonetheless, the utility of neuroscience and neuroimaging data in court is being increasingly realized as counsel, scientists, and practitioners work together to further establish its value in different contexts. Indeed, courts are often recommending, if not requiring, that neuroscience evidence be produced to support arguments being made, when these data are potentially informative (Catley and Claydon, 2016).

Understanding the many possible applications of neuroscience in the courts can help broaden this perspective and allay concerns that its arrival in legal proceedings is premature or imprudent. First, neuroscience evidence in the United States has almost never been an essential factor in determining *actus reus* – whether or not the accused actually committed the criminal act in question. However, for an interesting international example see State of Maharashtra v. Sharma[2]. and commentary by Gaudet (2011). Neuroscience evidence, in the U.S. and worldwide, is far more commonly determined to be relevant for assessing the competency of an individual to stand trial or in addressing the *mens rea* element of criminal liability, but for a more nuanced summary of these various contexts see Slobogin (2017). These latter contexts relating to *mens rea* often require thorough assessments of a defendant's mental health, which may reasonably include neuroimaging. Defenses built on this reasoning include pleas of insanity, which have evolved substantially over time (see section Our legal system is evolving, not static). Nonetheless, insanity pleas essentially argue that someone was so mentally impaired that s/he was not aware of their own actions, or able to decipher right from wrong. Still, a brain scan may only represent one element of a comprehensive clinical evaluation in such applications, or may be determined to be irrelevant given extant clinical evidence – that is, a brain scan is often not necessary for determining ones mental health status, but may provide supportive evidence. In other cases, neuroimaging may be essential (e.g., brain injury, degenerative disease, tumors). To be sure, insanity pleas overall constitute less than 1% of all felony trials, and their success ordinarily accompanies cases in which a defendant had previously been diagnosed with a mental illness (Kirschner and Galperin, 2001; Perlin, 2016).

As noted above, neuroscience data is also increasingly submitted in the sentencing phase of a trial (e.g., after guilt has already been determined), and evidentiary standards are somewhat more permissive and accommodating during these arguments. This is becoming more common in high stakes sentencing decisions, for instance, in capital cases when the convicted offender faces either the death penalty or life in prison (Miller, 2010). Neuroscience data may then be considered relevant when deciding whether one deserves the harshest possible punishment or a sentence reflecting intervening factors that can include mental health status. Indeed the relevance of mental health in some cases is considered so pertinent that a failure to introduce neuroscience data has been determinative of ineffective counsel and a violation of a defendant's constitutional right to fair representation (Koenig, 2016).

Additional applications of neuroscience arise in civil cases, which may require demonstrating extent of physical injury. Applications of brain imaging in this context include established documentation of gray and white matter injury with structural brain imaging, but may extend to novel applications that provide information on concussion and mild traumatic brain injury (Vergara et al., 2017), and application of functional imaging techniques addressing chronic pain (Wager et al., 2013). These later examples are emerging areas that need to be vetted further by the scientific community (Davis et al., 2017); however, their appeal and *potential* value is undeniable, underscoring the importance of lawyers and judges remaining in touch with advancing technology. In this way, it is imperative that legal counsel is adequately educated about the relevance and interpretation of neuroscience-based evidence that may aid the fair evaluation of each case.

It should be clear that, in any context, the tools of neuroscience are not subject to more lenient standards than other forms of evidence presented in legal arguments. That is, their probative value must be weighed against the potential for introducing a prejudicial impact or confusion among jurors. Commentaries often use this as a linchpin for their arguments, citing (limited) evidence that brain imaging evidence may mislead jurors and/or distract them from primary lines of reasoning (McCabe and Castel, 2008; Weisberg et al., 2008). Importantly, this evidence has been critically evaluated by others who have noted that these investigations did not present information in a context matching what juries typically encounter (Roskies et al., 2013). Other studies accounting for these factors have indicated no evidence that brain imaging carries any additional weight over and above verbal neuroscience-based testimony (Schweitzer and Saks, 2011; Schweitzer et al., 2011). Further, when evaluated in a legal context where cross-examination critically evaluates the relevance of information, MRI-based evidence is no more persuasive than other (non-neuroscience-based) evidence (McCabe et al., 2011). Finally, the role of the judge as a kind of gatekeeper for admissibility of evidence protects the system from more controversial applications of these tools. This has been demonstrated effectively time and again as courts have rejected the use of fMRI, for example, as a form of lie

[2] *State of Maharashtra v. Sharma*, C.C., No. 508/07, Pune, June 12, 2008 (India).

detection (US v. Semrau, 2010[3]; State v Gary Smith 2012[4].). This has persisted as there is not sufficient scientific consensus for these purposes, at present. As such, the use of fMRI in this context does not pass established Daubert standards of evidence.

The checks and balances built into the U.S. legal system have largely been effective in the face of expanding neuroscience evidence. It should be clear, however, that these safeguards are not intended to unilaterally prevent change in the legal system. Rather, they are intended to promote adaptive interpretation, reflecting normative standards that shift in step with increasing knowledge, advancing technology, and evolving public attitudes (see section "Our legal system is evolving, not static"). As technology continues to improve and new applications arise, it is essential for practitioners of the law to remain adequately informed in order to best serve their roles. The recognition of this imperative is increasingly evident in the resources and attention being devoted to these objectives in recent years (Jones et al., 2013). The MacArthur Foundation Law and Neuroscience project and the Research Network on Law and Neuroscience (MacArthur, 2019) represent large, multimillion dollar investments serving these needs. These efforts accompany many formal educational resources for judges and lawyers that specifically address topics of neuroscience (FJC, 2019), as well as ongoing development of many international conferences and academic societies devoted to increasing scholarship and improving communication within these integrative disciplines.

Our initial assertion in this commentary is intended to be uncontentious. Simply, there are many contexts in which the relevance of neuroscience data is already firmly established, and may in fact be essential for carrying out effective legal decision-making. Most of these applications are not new, but the breadth of their relevance has perhaps widened as their informative value improves in stride with progress in research and the technology itself. The relevance of neuroscience data in jurisprudence shows no evidence of diminishing in the coming years; therefore, we encourage an attitude of integration and motivated legal scholarship. The importance of this is clear even given the limited examples provided here, which leave out additional concerns regarding constitutional principles (Pizzetti, 2011), moral/ethical considerations, e.g., Pallarés-Dominguez and Gonzalez Esteban (2016); Shaw (2018), and Napier (2019), and emerging perspectives in international law (Spranger, 2012). Ongoing critical evaluation of the utility and limits of neuroscience will remain an essential component of this progress. Occasional dismissals of neuroscience's evolving relevance, in our view, are myopic and potentially dangerous. Criticisms on this order are usually intended to reinforce a static view that the law can continue to operate as effectively without neuroscience, simply because it has in the past. However, this attitude offers little guidance for the inevitable progress facing an assuredly dynamic field, which requires close evaluation of evolving technology and evidence. We therefore reinforce a perspective that the best way for the system to adapt to advancing technology is by improving education and resources available to legal professionals who are increasingly required to incorporate these data in their arguments.

OUR NORMATIVE UNDERSTANDING OF FREE WILL AND CULPABILITY IS CHANGING

Humans have been grappling with the concepts of free will and determinism (or *fatalism*) for most of recorded history (Hoefer, 2016); however, ancient notions of these concepts had little to do with the brain and neuroscience. Instead, philosophers and storytellers alike considered how much of our behavior was controlled by gods, the fates, or other supernatural external forces. Similarly, behavior that was attributable to our own motivations and decisions were not always nested in the brain. Aristotle for instance believed the brain mostly served to cool the blood. Rather, our motivated behavior has historically been attributed to something immaterial like a spirit, soul, or will. As physical sciences improved our understanding of neuromuscular junctions, neurotransmitters, and the role of the brain in organizing behavior based on prior experience, the role of a soul necessarily diminished. In his book *Soul Made Flesh*, Carl Zimmer develops a vivid history of neuroscience around the idea that advances in physiology, medicine, and psychology have incrementally narrowed the role of an immaterial soul as science has increasingly explained biological systems responsible for cognition and behavior (Zimmer, 2005).

In many ways, evolving perspectives about free will represent an extension of this trajectory. As neuroscience offers more detailed and predictive models accounting for human motivations, appetitive drives, and behavioral inhibition, extant descriptions of free will increasingly seem to grasp at something immaterial and elusive. This, somewhat covertly, promotes a paradigm incompatible with natural science, which progresses on a foundation that is fundamentally materialist, reductionist, and determinist in nature. As this represents a predictable extension of prior historical and philosophical progress, the questions neuroscience addresses on this topic are not new ones. However, neuroscience provides an increasingly tangible and convincing platform for demonstrating the limits, proximal antecedents, and illusions that support our subjective sense of free will. The relevance of this for promoting evolving attitudes in jurisprudence relate to how we, as a society, exercise normative judgments about agency, responsibility, and most importantly culpability. Here we illustrate how these attitudes are slowly shifting, and we emphasize the role neuroscience plays in influencing these standards.

Any treatment of how neuroscience has influenced our understanding of free will must address the studies of Benjamin Libet, and perhaps more importantly, contemporary extensions of this work. In the 1980s Libet published research demonstrating the precise timing of one's subjective perception of making a simple decision to freely move one's wrist in relation to other physiological events (Libet, 1985). The study essentially recorded three events: the movement of the wrist, the time the participant

[3] *US v. Semrau*, 693 F.3d 510 (6th Cir. 2012).

[4] *Smith v. State*, 32 A.3d 59, 423 Md. 573 (2011) (pretrial testimony).

"decided" to move their wrist, and neural activity surrounding these events. The neural activity indicative of preparations to move one's wrist was already known (since the 1960s), and is commonly referred to as the readiness potential (Kornhuber and Deecke, 1965). What was striking in Libet's experiments was the demonstration that this neural preparatory activity consistently preceded one's subjective sense of having decided to move, by about 350 ms. Preliminary interpretations of these outcomes suggested that neural activity preceding the decision-point constituted evidence of a deterministic process that had already begun, prior to our subjective awareness of it, undermining conventional notions of agency or free will more generally (Libet, 1999). These initial conclusions have been rightfully debated for decades, while others have more quietly continued to improve and expand on these methods.

More recent extensions of this work have included the application of machine learning algorithms to accurately predict subjects' movements before they decide to move. This has been carried out using intracranial, intracellular recordings within the supplementary motor cortex (Fried et al., 2011). Functional MRI recordings measuring patterns of neural activity across the whole brain have also reliably predicted which of two buttons someone will press up to 7 s before they indicate they've decided (Soon et al., 2008). However, other exciting research demonstrates that these behaviors are not determined in such a simplistic way; but rather, they remain influenced by parallel cognitive systems. Executive control systems can essentially veto an intended movement, if given as little as 200 ms warning (Schultze-Kraft et al., 2016). That is, a 'stop' signal triggered by a real-time prediction of one's intended movement is sufficient to allow inhibition and eliminate that movement, provided it is delivered at least 200 ms prior to the execution of the event.

Demonstrations like these provide concrete evidence that our decisions and motivations are accompanied by many parallel neural mechanisms that occupy a dimension beyond our conscious, deliberative processes of reasoning. Measuring the corresponding neural activity provides tangible, proximal measurements of these processes, but neuroscience is not the only context in which we are aware of such unconscious influences on behavior. Freud may have been the first to draw public attention to the prominent role of subconscious influences on our otherwise rational behavior (Freud, 1913). More recently, studies of decision-making in contexts ranging from economics to moral deliberation have made it clear that our choices are strongly guided by implicit emotional influences that often deviate from rational optimization, and the narratives we construct around our decisions are often architected in a *post hoc* manner (Haidt, 2001; Lerner et al., 2015). Finally, we are increasingly aware of the predictable consequences of many remote influences that we have no individual control over. These include our genetic makeup (Brunner et al., 1993; Mason and Frick, 1994), early rearing environments (Kaplow and Widom, 2007; Mulvaney and Mebert, 2007), and complex social systems (Yoshikawa et al., 2012; Javanbakht et al., 2015). These factors all bias our cognition and behavior in predictable ways and their influence impinges on neural systems that guide our behavior directly, in ways that are largely inaccessible to our moment-to-moment conscious, deliberative processes. Better understanding of these influences has, in some ways, also changed the way we reason about culpability.

Challengers to the role of neuroscience in legal contexts will often argue that claims of functional impairments based in neuroscience contribute little to our normative judgments about culpability. This is based, in part, on the rationale that innumerable others with similar impairments have undoubtedly not committed similar crimes, and so the impairment (by itself) is insufficient to predestine the crime (Morse, 2006; Mayberg, 2010). This rebuttal fails to recognize that deterministic influences rarely operate in isolation, and our *normative* judgments ordinarily consider multiple factors and contextual circumstances (Freedman and Zaami, 2019). Further, the deterministic limits of isolated factors on criminal behavior are not uniquely reserved for neuroscientific considerations. This same argument, for instance, fails to undermine the relevance of something like faulty brakes influencing our normative judgments about a fatal car accident, given that faulty brakes only sometimes lead to fatalities, see also (Zeki et al., 2004). This perspective also misses a somewhat more overarching role that neuroscience plays in shifting normative judgments about culpability. That is, neuroscience can help shift our judgments by simply grounding facts about psychological differences in a physical realm, underscoring their contextual relevance among many forms of physical evidence.

If the processes of motivation and decision-making are seen only as imponderable mysteries, inaccessible to reductionistic science, then we are constrained by limited insight into the origins of behaviors we ostensibly wish to diminish in society. We are further bound to make more simplistic normative judgments based on right and wrong, and our interventions will be more unidimensionally focused on reactive punishment. Conversely, integrating deterministic perspectives in explaining behaviors society condemns doesn't prevent us from using punishment as a deterrent, but only highlights additional points of leverage useful for applying more proactive interventions as an added method of diminishing unwanted behavior, see also Eagleman (2011b) and Slobogin (2011, 2017).

Another interesting context from which to observe this evolving landscape is to consider relatively common forms of pathology that impinge on our ability to choose and behave freely. Fitting examples include obsessive-compulsive disorders and addictions. In both cases, individuals can be said to lose some control over behavior that, in healthy individuals, is attributed to ordinary volitional processes. Normally, washing our hands, going over a mental list, or enjoying a beer are all considered among our ordinary, voluntary, healthy behaviors. Under pathological conditions, however, compulsions to engage in these or other behaviors encroach on (and supersede) other normal motivations. Daily goals, long-term ambitions, and explicit objectives may be at odds with increasingly intrusive thoughts and behaviors that an individual has limited control over. An individual may fully understand, anticipate, and wish to avoid the consequences of certain maladaptive behaviors, while still succumbing to well-worn patterns leading to the undesired behavior. Common understanding about the pathophysiology of

these disorders has altered how we address these issues both clinically and interpersonally.

The current accepted model of addiction promoted by the National Institutes of Health is that of a brain disorder instantiated in motivational and inhibitory systems, brought on by exposure to substances that pharmacologically impose lasting physiological changes on these systems (NIDA, 2019). Like other diseases, genetic vulnerabilities, environmental exposures, and variability in physiology all promote individual differences in susceptibility to addiction. Unlike many other diseases, the observable symptoms are almost entirely behavioral. Moreover, these behaviors are often categorically illegal and punishable by law (in the case of illicit drug use), but may also be viewed under a moral lens as a transgression against more virtuous decision making. What makes this acutely relevant to discussions of neurolaw is the nature of arguments for and against the disease model of addiction, and how they reflect philosophical discourse on free will, neuro-determinism, and culpability. Opposition to the disease model can be easily found in publications such as Heyman's *Addiction: A Disorder of Choice* (Heyman, 2009), Schaler's *Addiction is a Choice* (Schaler, 2000), and Satel and Lilienfeld's *Addiction and the Brain-Disease Fallacy* (Satel and Lilienfeld, 2014). These arguments make rhetorical appeals to the primary role of *choice*, *agency*, *volition*, and *self-control*. In doing so, they tacitly place limits on reductionist approaches that examine supportive physiological processes that govern our choices. These arguments seem rooted in the fundamental conservation of free will as something irreducible, and impervious to reductionist, deterministic paradigms.

By contrast, neuroscientific research nested in the disease model of addiction studies elements of motivated behavior in simpler parts, examining individual variability across these dimensions. For example, this research examines shifts in valuation (e.g., the motivational weight of pharmacological reinforcers over natural reinforcers) along with the weakening of inhibitory control (Goldstein and Volkow, 2002). These approaches also examine transitions between behavior governed primarily by executive control systems and behavior carried out by networks governing compulsive, automatic actions (Kalivas and O'Brien, 2008). As such, the "disease" aspect of addiction is more fundamentally rooted in the physiological systems that govern our choices and behaviors, rather than in the complex behavioral symptom of drug-taking *per se*. Adopting this perspective requires a reductionist and determinist paradigm for informing our free will. While not universally accepted (and perhaps still requiring semantic refinement), the progressive contributions of the disease model of addiction include a better understanding of biological influences that culminate in our motivated behavior. Progress on this front further serves to dilute a predominantly moralistic attitude toward addiction that may motivate primarily punitive actions intended to address a very legitimate societal problem. This contribution will feature heavily in our ongoing assessment of the relevance of neuroscience in an evolving criminal justice system.

These arguments are familiar in the context of debates about the nature of free will and responsibility. While our intuitions may still demand the preservation of a concept like free will and agency in our behavior (Nahmias et al., 2005; Nichols, 2011), it has become increasingly necessary to clarify what aspects of our thoughts and behavior remain free, to what extent they are free, and (perhaps most important for our purposes here) what the relevance of this is for our judgments about how to intervene to address pragmatic social needs. After all, moral responsibility is more abstract and partially removed from the practical considerations of punishment and intervention in our justice system. As it turns out, laypeople's judgments about these topics are not always internally consistent, often reflecting shifting attitudes when considered in abstract terms vs. concrete examples. For instance, when considering theoretical arguments, people are more likely to maintain that determinism undermines basic moral responsibility; when considering concrete episodic scenarios, we are more likely to affirm basic accountability for our actions responsibility (Nichols, 2011).

Using the disease model of addiction as an example, opponents do not deny the evidence of biological changes in motivational systems that account for changes in behavior. However, opponents still cling to the relevance of individual agency, free will, and decision making, ostensibly apart from their biological influences, perhaps only because this reaffirms our most basic intuitions about choice, consequences, and our ability to change (Feldman et al., 2014). This veneration of free will over the biological systems that govern choice may have counterproductive consequences, however. The best methods for intervention arguably improve by understanding the biological systems governing our choices and motivated behavior, particularly in the context of maladaptive behaviors involving substances that impinge directly on these systems. It is the context of intervention that becomes highly relevant for our consideration of the ongoing role of neuroscience in the future of jurisprudence. The influence of neuroscience on these concerns is already evident in a number of contexts discussed in the next section, and it has the potential to continue to improve our practical management of an imperfect but adaptable criminal justice system.

OUR LEGAL SYSTEM IS EVOLVING, NOT STATIC

The relevance neuroscience has in our current justice system is already firmly established in several contexts outlined in Section "Neuroscience Has Firmly Established Its Place in Jurisprudence." The way neuroscience is promoting progressive changes in our justice system is also evident in a number of ways we address here. We can use recent examples of these changes to help anticipate the ongoing evolution of jurisprudence as informed by advancing neuroscience. Importantly, we reiterate our position that the influence of neuroscience has relatively less to do with any perceived exculpatory extensions of a purely deterministic universe (*my brain made me do it*), and is more practically relevant for the way we interpret concepts like "justice" and the role of the justice system in promoting a safe, functioning society. Shifting normative attitudes on this front influence how we choose to intervene and hold

people accountable for their actions. Neuroscience, after all, has improved our general understanding of motivated human behavior and myriad deterministic influences that converge to promote maladaptive, antisocial behavior. Where there is improved understanding of these influences, we will be better equipped to introduce improved strategies at remedying systemic problems contributing to the behaviors and societal problems we aim to diminish.

Conservative appeals to traditional applications of jurisprudence regularly make the claim that neuroscience need not change anything about the way we interpret legal responsibility. This is true in one sense: if our only motivation is preservation of the *status quo*. In an article previously published in this series, *Criminal Responsibility and Neuroscience: No Revolution Yet*, Bigenwald and Chambon (Chambon and Bigenwald, 2019) establish that no revolution is necessary for us to continue applying the same normative framework of responsibility that the legal system has always operated on. Several arguments are presented emphasizing the primacy of our intuitions about agency for assigning criminal responsibility. That is, even the reality of a purely deterministic universe does not negate criminal responsibility, which in their view, exists as a mostly pragmatic concept independent of free will. This is true only in that our legal system certainly employs a number of arbitrary rules in order to remain serviceable. It is no great leap in understanding to suggest that it simply operates 'as if' we are responsible agents. Our objection on this matter is that this reality will become increasingly dissatisfying, even from a normative perspective, as general knowledge increases, providing more insight into the boundaries and limitations of our own agency. Fortunately, the present series of articles is under no obligation to preserve the *status quo*; but rather, it challenges us to describe how the legal system might be practically changed by discoveries in neuroscience. We therefore submit that these changes may be less visible in the ways we interpret and enforce the law, and more visible in the ways we punish violators of our laws and adapt as a society to preserve (or advance) our most pressing goals.

As Bigenwald and Chambon point out, *responsibility* has many possible meanings. A tree can certainly be responsible for falling on a wire and causing a power failure, even though it has no real agency. Calling a tree responsible for these consequences doesn't violate any of our intuitions about agency and its value. Calling a tree "guilty" for this, however, feels odd (violates our intuitions), just as wishing to implement retributive harm on the tree would seem senseless. This illustration emphasizes a division between practical considerations of responsibility and the attribution of a kind of value judgment about the tree's actions, and how they align with normative moral values. Even as we are keenly interested in (also) preventing other trees from falling (deterrence), one of the key roles of our justice system remains a punitive one, and this features heavily in how harshly we decide to punish. Our intuitions about agency, free will, and moral judgment play a much larger role in our instinct to punish the guilty. Where neuroscience may play its most significant role is in the space between legal determinations and implementing corrective measures that benefit society. Here there remains a great deal of room for improving strategies aimed at protecting and benefiting society on a large scale. These changes in normative attitudes are evident in the evolving standards we use in legal sentencing and the ways we continue to evaluate the relative efficacy of various punitive strategies.

As noted, the criminal justice system in the United States serves many functions beyond a punitive one. We rely on it to deter flagrant abuses of the law, to protect society at large from the most dangerous individuals, and (ostensibly) to help intervene and rehabilitate those who violate the expectations of their social systems. The current implementation of this system, however, has been heavily biased toward retributivist deterrence strategies, which have demonstrated their own limitations over several decades (Frost, 2006). Indeed, they have contributed, in part, to the highest incarceration rates, per capita, in the entire world. Public attitudes play an overt role in this as the Supreme Court has endorsed that public desire for retribution is a legitimate basis for establishing harsh, punitive judgments up to and including capital punishment (Gregg v. Georgia, 1976[5]).

Initial steps in adopting more effective strategies may be fostered by increasing numbers of people reconsidering the implicit relevance and meaning of concepts like *free will* for achieving societies' goals. As the meaning of this concept evolves and our understanding of behavior integrates more deterministic features, we are less compelled to frame maladaptive, antisocial actions within paradigms that embrace elusive immaterial origins (like evil, for instance) (Grasmick et al., 1992; Unnever et al., 2005). Rather, we are better equipped to recognize the influence of pathology, environment, and acquired maladaptive cognitive strategies in promoting antisocial behavior, where the levers of justice have considerably more remedial influence. After all, pathology is a more tractable problem than is evil. Responding to antisocial behavior, then, becomes a more pragmatic issue, and more progressive strategies aimed at addressing objective moderators of such behavior can be readily explored. In this way, even slow shifts in normative judgments are highly relevant to the way we assess culpability as a society, and the degree to which we view punitive measures as achieving their intended purpose as a remedy against undesired, antisocial behavior.

Neuroscience ultimately provides a useful platform for advocating new strategies of social management, where old strategies have perhaps proven ineffective or inefficient. New strategies may be less oriented toward retribution, *per se*, and more driven by practical concerns serving society with more efficient and productive solutions. Such strategies may, for instance, be aimed at better serving the mental health and social needs of those who come in contact with the justice system, reducing long-term incarceration rates for low risk offenders, and reducing recidivism by improving rehabilitative and reintegration efforts. In the worst scenarios, where rehabilitative interventions may not be a realistic goal, neuroscience also provides a platform for improving our predictions of future dangerousness (Aharoni et al., 2013; Steele et al., 2015; Kiehl et al., 2018). Such strategies may be integrated for making better decisions about those who need to be removed from society

[5] *Gregg v. Georgia*, 428 U.S. 153, 96 S. Ct. 2909, 49 L. Ed. 2d 859 (1976).

permanently. Before addressing this further, it will be important to consider a few examples for how advancing science has already changed the way we think about culpability and make decisions about interventions and punishment as a society.

Limits on Capital Punishment

Torture and execution have been legally sanctioned forms of punishment since at least the 18th century BCE, as it is indicated in the Code of Hammurabi (c.1750 BCE) for such crimes as burglary, adultery, making false accusations, and poor construction of a house (Harper, 1904). Even within the history of the United States, the use of capital punishment has changed considerably, formerly implemented in cases of burglary, counterfeiting, and treason among others (Randa, 1997). Beginning with the adoption of bans on cruel and unusual punishment[6]., modern societies (including the U.S.) have gradually changed their views on behavior deserving the harshest penalties, limiting its application for the most egregious crimes and even further to individuals most deserving of harsh punishment. Determining who deserves the harshest punishments has a great deal to do with our perceptions of their intentions, malice, and reasonable expectations of self-control. As we will see, these judgments also incorporate the relative utility of the punishment for fulfilling an intended punitive role. The relevance of neuroscience in drawing conclusions about these issues increases as their evaluation increasingly incorporate reductionist, determinist, biological perspectives of motivated behavior.

Prominent examples of this have come in the form of supreme court decisions accompanying restricted applications of the death penalty. For instance, Atkins v. Virginia[7]. ruled to prevent the execution of those with severe intellectual disabilities, citing "evolving standards of decency that mark the progress of a mature society." In cases like these, these *evolving standards* refer more specifically to what the court witnesses as a consensus among other jurisdictions and the way they have tended to interpret and enforce the law in recent history. Many states, for instance, had previously outlawed the execution of those with severe intellectual disabilities prior to these proceedings. Among the topics discussed in the formal ruling is the sentiment that those with reduced intellectual capacity have limitations in their adaptive functioning, reasoning, communication, and understanding of events around them and the actions of others. Thus, leveraging the most severe of punishments fails to align with the practical concerns of retribution and deterrence.

Similarly, Roper v. Simmons (2005)[8]. abolished capital punishment of juveniles citing similar "evolving standards" and an emerging consensus among other jurisdictions. In this case, however, the decision was also influenced in part by neuroscience research (including fMRI evidence) presented in an amicus brief by the American Psychological Association, suggesting that psychological deficits germane to adolescence (developmental limitations) make young people more prone to impulsive behavior and less capable of the highest order decision-making we ordinarily attribute to adults. That is, opinions informed by progress in neuroscience suggesting a limited capacity for behavioral control are influential for evaluating an individual's culpability (i.e., how harsh a punishment is justified). Implicit in the developmental perspective applied is an acknowledgment of the capacity for ongoing change. MRI evidence was also presented (in amicus brief) for consideration in Graham v. Florida (2011)[9]. which determined it unconstitutional to sentence juveniles to life without parole for crimes not involving homicide.

These decisions can be fully reconciled with normative attitudes about responsibility and determinism. Despite being fully responsible for their behavior, biological limitations on individuals' executive functioning and inherent capacity for change play a prominent role in our consideration of how harsh their punishments ought to be. The Court's decision in Roper v. Simmons affirmed that juveniles have less culpability due to their immature development, making them less deserving of the harshest punishments. These decisions do not imply that, as a society, we are any less interested in protecting ourselves from dangerous people or ensuring the safety of free citizens. What is confirmed in these decisions is a relative diminution in our motivation to levy harsh retributivist judgments in contexts where we recognize deterministic limitations in individual culpability. This, of course, opens the door to consider how we judge those with other biological limitations in cognitive functioning, or those disadvantaged in other ways.

The Insanity Defense

Our collective understanding of culpability has almost always included provisions for certain disadvantages. A clear illustration of this endures in the limitations on culpability levied against those with serious mental disorders. This has been a common feature of many ancient legal systems and customs, including elements of Roman law which were carried forward in pre-Norman England (Walker, 1985). For instance, it was at times customary for juries to find insane criminals guilty, but refer them to the king for subsequent pardoning. More contemporary applications of these provisions give juries specific guidelines for applying these judgments directly. The *M'Naghten Rule*, for instance, formalized a set of conditions in English law that could be applied more consistently following a controversial acquittal. In 1843, Daniel M'Naghten suffered paranoid delusions and murdered a civil servant, mistaking him for the English Prime Minister. He was acquitted for murder based on substantial evidence that he was mentally ill, and he was forcibly committed to an asylum, where he spent the rest of his life. Despite the very real limits placed on his freedom, the ensuing public dissent following a *not-guilty* verdict (and official condemnation of the verdict by the queen) compelled establishing a set of explicit requirements for instantiating criminal insanity. These guidelines, in some adapted form, are still prevalent in many jurisdictions across the world today. They essentially require (for an insanity defense) that a defendant be so mentally

[6] U.S. Const. amend. VIII.

[7] *Atkins v. Virginia*, 536 U.S. 304, 122 S. Ct. 2242, 153 L. Ed. 2d 335 (2002).

[8] *Roper v. Simmons*, 543 U.S. 551, 125 S. Ct. 1183, 161 L. Ed. 2d 1 (2005).

[9] *Graham v. Florida*, 560 U.S. 48, 130 S. Ct. 2011, 176 L. Ed. 2d 825 (2010).

impaired as to not know what they are doing and/or not know right from wrong.

Following in step with the very impetus for M'Naghten, many subsequent adaptations and amendments to these rules have been applied, usually following controversial rulings. As a result, several variations and alternative defenses have been enacted in state and federal jurisdictions. These either amend the essential language of M'Naghten used to describe what constitutes insanity, or they shift the burden of proof in important ways. For instance, in Parsons v. State of Alabama (1887)[10]., an appeal was made following the controversial conviction of Nancy Parsons who killed her husband under the delusion that he had cast an evil spell on her. The court established a provision for instances in which a defendant may be deemed insane, despite knowing right from wrong. The ruling described instances where a disease has "destroyed the defendant's free will" and became known as the *Irresistible Impulse* defense. Other important developments have included modifications that specifically exclude antisocial personality disorder from an exculpatory mental illness, since its symptoms are primarily manifest through repeated criminal conduct (American Law Institute Model Penal Code, 1962)[11]. There has also been a formal shift of responsibility from the prosecution – proving beyond a reasonable doubt that a defendant was sane – to the defense, which must prove (by preponderance of evidence) that the defendant was insane (Insanity Defense Reform Act; IDRA, 1984)[12]. A lengthy, stand-alone review would be necessary to adequately review the many important modifications that have been made to these rules over time and across many jurisdictions; however, an overarching pattern is apparent in this complex history. Through many shifts of language and interpretations, we continually re-affirm the preservation of limitations on culpability for those impaired by mental illness. We also betray the cognitive dissonance this instills against the backdrop of our most basic retributive motivations, and our sensitivities to potential abuses of these provisions.

As noted above, many of these changes come on the heels of controversial, high-profile cases. Consider for instance the trial of Dan White for the murder of San Francisco Mayor George Moscone and city supervisor Harvey Milk. Despite substantial evidence of premeditation and malice in the killings, White was ultimately convicted of voluntary manslaughter rather than first-degree murder, and served only 5 years in prison. This outcome was aided by what is still disparagingly referred to as "The Twinkie Defense." To this day, popular retellings of this case often reinforce a narrative that White's defense asserted his behavior was the result of eating sugary snacks, including Twinkies. In reality, psychiatrists testified that White suffered from major depression and had diminished capacity for controlling his behavior due to this pathology. An incidental detail of his diminished capacity included recent poor dietary habits, despite having been extremely health conscious all his life. The public outcry following his alarmingly lenient sentence was instrumental in abolishing the "diminished capacity" defense in California. The relevance of this trial for our present arguments is not so much to draw attention to the trial and defense, but rather what happened afterward. Following a defense which hinged on severe depression, White served a relatively lenient sentence at Soledad State Prison in California (not a secure hospital, or institute known for therapeutic intervention). Two years after his release from prison, Dan White committed suicide.

These events raise many interesting questions from a legal, psychological, and philosophical perspective. Did White ultimately get what he deserved? Did those seeking harsher retributive action find some gratification in his suicide, or is it inherently less satisfying that White's death was not carried out in a punitive context? Was White's case a greater failure in the basic judicial process of determining culpability, or more of a failure in enforcing effective interventions following a determination of his mental illness? Those like White, committing serious offenses (e.g., homicide) in the throes of mental illness, are typically forcibly committed to secure institutions with some focused psychiatric capacity, and do not generally go free in such a short time. From a utilitarian perspective, this form of intervention seems reasonable. It serves the role of protecting society from dangerous people and arguably remains a visible deterrent, while coupling offenders' containment with therapeutic and/or rehabilitative attention. Where this strategy fails is perhaps only limited in satisfying an instinctual urge to serve harsh retributive actions against those that have harmed us personally and/or violated our most sacred moral values (Grasmick et al., 1992; Unnever et al., 2005).

It matters that this trial has largely been enshrined as a miscarriage of justice, but probably for many of the wrong reasons (as evinced by history's retelling of the "Twinkie Defense"). In some ways eradication of the diminished capacity defense serves as a scapegoat that only distracts us from more fundamental issues in our society and our justice system that are slow to change. Essentially, we still struggle to balance our shifting attitudes of culpability against a stubborn instinct to enact harsh retributive penalties in cases of egregious tragedy. After all, various provisions for mental illness in sentencing still exist in virtually all jurisdictions. Despite fine tuning the language of these rules, we (as a society) have steadfastly acknowledged that criminal actions occurring due to factors outside the ordinary limits of one's control deserve more leniency or a categorically different form of intervention than simple retribution. In order to see the potential benefit of progressive changes in jurisprudence, however, our corrections systems and forensic psychiatric facilities need to be equipped with the tools to enact these changes in ways that demonstrate more satisfying results.

How the Legal System May (Continue to) Change

The contexts described above illustrate that our normative views of culpability have never been static, but continually adapt to

[10] *Parson v. State*, 81 Ala. 577, 2 So. 854, 2 So. 2d 854 (1887).

[11] American Law Institute: Model Penal Code. Philadelphia: ALI, (1962) Ref. 11, § 4.01

[12] Insanity Defense Reform Act ("IDRA"), 18 U.S.C. § 17(b) (1984).

evolving standards ushered in by a more refined understanding of human behavior and the boundaries of our own agency. Neuroscience, surely, does not make this progress simpler. On the contrary, as our understanding of biology's role in promoting pathology and maladaptive behavior increases, this encourages a far more nuanced interpretation of culpability in the face of various advantages and disadvantages. Shifting attitudes on this front have fostered an expansion of contexts that we harbor special provisions for in the law. However, rather than promoting overly exculpatory attitudes (a kind of *straw man* common in arguments diminishing the role of neuroscience in law), these shifts have largely required changes only in the way we intervene and balance corrective and/or punitive measures in such contexts. We suggest it is reasonable to expect these trends to continue as retributive goals are softened and we aim to integrate more practical solutions for addressing criminogenic needs, improving reintegration, and reducing recidivism.

Using offender age as a model for such a transition, the criminal justice system has recently made provisions to limit harsh sentencing (capital punishment/life in prison) of juvenile offenders in most circumstances, but various jurisdictions still apply somewhat arbitrary rules about the cut-offs for these provisions. Certainly young offenders are not all equal from a neurodevelopmental perspective. So does it make sense to apply a bright line rule allowing capital punishment on someone's 18th birthday? Neuroscience may continue to inform this perspective, expanding a more nuanced evaluation. Recent research, in fact, has demonstrated that a brain-derived measure of gray matter related to age is a better predictor of future antisocial behavior than is chronological age (Kiehl et al., 2018). As such, it may be a more pertinent question to consider the relative advantages in development and mental health with which one is equipped before deciding whether they deserve our harshest punishments. Trends in this direction are encouraged by the bifurcation of guilt and sentencing phases of some criminal trials. Another perspective to consider is what effort and resources are justified in the aim of preserving and enacting capital punishment as an extreme punitive measure for rare circumstances. Studies on the deterrence effects of the death penalty are equivocal at best (Weisberg, 2005), and economic scrutiny suggests that we may be incapable of enacting the death penalty in any reasonably efficient manner such that it serves its intended purposes (Aviram and Newby, 2013). Despite our enduring retributivist instincts, we may eventually decide that abolition of the death penalty represents a more practical solution, obviating some of our more difficult choices when it comes to punishment. But capital punishment is not the only context within which shifting attitudes may promote more pragmatic strategies.

In the case of less severe sentences, we (as a society) have demonstrated more amenability to the potential value of remedial approaches and possible re-integration of young offenders. Certainly, more aggressive treatment strategies integrating contemporary cognitive-behavioral approaches for improving long-range outcomes have proven both successful and cost-effective (Caldwell and Van Rybroek, 2013). Consider for instance progressive treatment programs being instituted at the Mendota Juvenile Treatment Center among high-risk young offenders (Caldwell and Van Rybroek, 2005; Caldwell et al., 2006a). Analyses have indicated better outcomes and relatively less economic burden on society by enacting these aggressive treatment strategies (Caldwell et al., 2006b). To be sure, there will always be those who are resistant to our best available treatments at any given time, and unable to return safely to free society. However, this further reinforces the value of pursuing new and better strategies informed by ongoing research addressing the origins, development, and maintenance of maladaptive, antisocial behavior.

In assessing how neuroscience may continue to inform judicial decision-making in the future, many possibilities arise. Could brain scans that objectively quantify one's neurodevelopment or functional capacity eventually be used to determine whether one is tried as an adult or juvenile? Could predictive models determine whether one is amenable to therapeutic attention or is likely to remain resistant to available rehabilitative efforts? Could neuroscientific measures reveal specific risk factors for re-offending that are not evident on standard psychiatric assessments? These are difficult questions indeed, and ones that we do not yet have answers for. We simply argue that to ignore them or to undermine their potential value only to reinforce the *status quo* seems myopic and overtly servile toward an imperfect system. As neuroscience ushers in a more complete bio-psycho-social understanding of maladaptive behavior, and as ongoing incarceration strategies become unsustainable, our prediction is that we will be forced to consider alternative approaches that serve public interest in more pragmatic ways. This will involve wider application of therapeutic, rehabilitative approaches and more aggressive therapeutic and reintegration strategies that reduce the likelihood for recidivism. Such applications may be particularly effective among young offenders (Glenn, 2019). This may also include better risk assessment in making decisions about sentencing and parole. Major advances on these fronts may only require us to first suspend our most basic retributivist instincts when addressing social problems, and remain open minded about the potential for more prudent strategies. Neuroscience doesn't fill these roles on its own, but it provides a platform for advancing each of these goals through empirical research and improved knowledge.

AUTHOR CONTRIBUTIONS

NA developed the primary theses and arguments presented in this review. KK provided additional insights, commentary, and editorial remarks.

ACKNOWLEDGMENTS

Commentaries contained in this review reflect the thoughts and opinions of the authors only, and do not reflect official strategies or priorities of The Mind Research Network, funding bodies, state/federal organizations, or other supporters of our research.

We would like to thank the staff and administration of the New Mexico Corrections Department and Wisconsin Department of Corrections for their continued cooperation and support with ongoing research at the Mind Research Network. We also thank the volunteers participating in research and the research staff that make our research possible.

REFERENCES

Aharoni, E., Vincent, G. M., Harenski, C. L., Calhoun, V. D., Sinnott-Armstrong, W., Gazzaniga, M. S., et al. (2013). Neuroprediction of future rearrest. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6223–6228. doi: 10.1073/pnas.1219302110

Aviram, H., and Newby, R. (2013). Death row economics: the rise of fiscally prudent anti-death penalty activism. *Crim. Just.* 28, 33–40.

Brown, T., and Murphy, E. (2010). Through a scanner darkly. *Stanford Law Rev.* 62, 1119–1208.

Brunner, H. G., Nelen, M., Breakefield, X., Ropers, H., and Van Oost, B. (1993). Abnormal behavior associated with a point mutation in the structural gene for monoamine oxidase A. *Science* 262, 578–580. doi: 10.1126/science.8211186

Burns, K., and Bechara, A. (2007). Decision making and free will: a neuroscience perspective. *Behav. Sci. Law* 25, 263–280. doi: 10.1708/2631.27049

Caldwell, M., Skeem, J., Salekin, R., and Van Rybroek, G. (2006a). Treatment response of adolescent offenders with psychopathy features: a 2-year follow-up. *Crim. Justice Behav.* 33, 571–596. doi: 10.1177/0093854806288176

Caldwell, M. F., and Van Rybroek, G. (2013). Effective treatment programs for violent adolescents: programmatic challenges and promising features. *Aggress. Violent Behav.* 18, 571–578. doi: 10.1016/j.avb.2013.06.004

Caldwell, M. F., and Van Rybroek, G. J. (2005). Reducing violence in serious juvenile offenders using intensive treatment. *Int. J. Law Psychiatry* 28, 622–636. doi: 10.1016/j.ijlp.2004.07.001

Caldwell, M. F., Vitacco, M., and Van Rybroek, G. J. (2006b). Are violent delinquents worth treating? A cost–benefit analysis. *J. Res. Crime Delinq.* 43, 148–168. doi: 10.1177/0022427805280053

Catley, P., and Claydon, L. (2016). The use of neuroscientific evidence in the courtroom by those accused of criminal offenses in England and Wales. *J. Law Biosci.* 2, 510–549.

Cave, S. (2016). *There's no Such Thing as Free Will: But We're Better Off Believing in it Anyway. The Atlantic*. Available online at: www.theatlantic.com/magazine/archive/2016/06/theresno-such-thing-as-free-will/480750/?utm_source=atlfb (accessed November 15, 2019).

Chambon, V., and Bigenwald, A. (2019). Criminal responsibility and neuroscience: no revolution yet. *Front. Psychol.* 10:1406. doi: 10.3389/fpsyg.2019.01406

Davis, K. D., Flor, H., Greely, H. T., Iannetti, G. D., Mackey, S., Ploner, M., et al. (2017). Brain imaging tests for chronic pain: medical, legal and ethical issues and recommendations. *Nat. Rev. Neurol.* 13, 624–638. doi: 10.1038/nrneurol.2017.122

De Kogel, C., and Westgeest, E. (2016). Neuroscientific and behavioral genetic information in criminal cases in the Netherlands. *J. Law Biosci.* 2, 580–605.

Eagleman, D. (2011a). *The Brain on Trial. The Atlantic.* Available online at: https://www.theatlantic.com/magazine/archive/2011/07/the-brain-on-trial/308520/ (accessed November 15, 2019).

Eagleman, D. (2011b). The brain on trial. *Atlantic* 7, 112–123.

Farahany, N. A. (2016). Neuroscience and behavioral genetics in US criminal law: an empirical analysis. *J. Law Biosci.* 2, 485–509.

Feldman, G., Baumeister, R. F., and Wong, K. F. E. (2014). Free will is about choosing: the link between choice and the belief in free will. *J. Exp. Soc. Psychol.* 55, 239–245. doi: 10.1016/j.jesp.2014.07.012

FJC (2019). *Educational Resources, Neuroscience* [Online]. Washington, DC: Federal Judicial Center.

Freedman, D., and Zaami, S. (2019). Neuroscience and mental state issues in forensic assessment. *Int. J. Law Psychiatry* 65:101437. doi: 10.1016/j.ijlp.2019.03.006

Freud, S. (1913). *The Interpretation of Dreams*. New York, NY: The Macmillan Company.

Fried, I., Mukamel, R., and Kreiman, G. (2011). Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron* 69, 548–562. doi: 10.1016/j.neuron.2010.11.045

Frost, N. A. (2006). *Punitive State: Crime, Punishment, and Imprisonment Across the United States*. El Paso, TX: LFB Scholarly Publishing.

Gaudet, L. M. (2011). Brain fingerprinting, scientific evidence, and Daubert: a cautionary lesson from India. *Jurimetrics* 51, 293–318.

Glenn, A. L. (2019). Using biological factors to individualize interventions for youth with conduct problems: current state and ethical issues. *Int. J. Law Psychiatry* 65:101348. doi: 10.1016/j.ijlp.2018.04.008

Goldstein, R. Z., and Volkow, N. D. (2002). Drug addiction and its underlying neurobiological basis: neuroimaging evidence for the involvement of the frontal cortex. *Am. J. Psychiatry* 159, 1642–1652. doi: 10.1176/appi.ajp.159.10.1642

Gonzalez, R. (2017). *How Criminal Courts are Putting Brains–Not People–on Trial. Wired*. Available online at: https://www.wired.com/story/how-criminal-courts-are-putting-brains-not-people-on-trial/ (accessed November 15, 2019).

Grasmick, H. G., Davenport, E., Chamlin, M. B., and Bursik, R. J. Jr. (1992). Protestant fundamentalism and the retributive doctrine of punishment. *Criminology* 30, 21–46. doi: 10.1111/j.1745-9125.1992.tb01092.x

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychol. Rev.* 108, 814–834. doi: 10.1037/0033-295x.108.4.814

Harper, R. F. (1904). *The Code of Hammurabi King of Babylon*. Chicago, IL: The University of Chicago Press.

Harris, S. (2012). *Free Will*. New York, NY: The New York Times.

Heyman, G. M. (2009). *Addiction: A disorder of Choice*. Cambridge, MA: Harvard University Press.

Hoefer, C. (2016). "Causal Determinism," in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Stanford, CA: Stanford University).

Javanbakht, A., King, A. P., Evans, G. W., Swain, J. E., Angstadt, M., Phan, K. L., et al. (2015). Childhood poverty predicts adult amygdala and frontal activity and connectivity in response to emotional faces. *Front. Behav. Neurosci.* 9:154. doi: 10.3389/fnbeh.2015.00154

Jones, O. D., Marois, R., Farah, M. J., and Greely, H. T. (2013). Law and neuroscience. *J. Neurosci.* 33, 17624–17630.

Kalivas, P. W., and O'Brien, C. (2008). Drug addiction as a pathology of staged neuroplasticity. *Neuropsychopharmacology* 33, 166–180. doi: 10.1038/sj.npp.1301564

Kaplow, J. B., and Widom, C. S. (2007). Age of onset of child maltreatment predicts long-term mental health outcomes. *J. Abnorm. Psychol.* 116, 176–187. doi: 10.1037/0021-843x.116.1.176

Kiehl, K. A., Anderson, N. E., Aharoni, E., Maurer, J. M., Harenski, K. A., Rao, V., et al. (2018). Age of gray matters: neuroprediction of recidivism. *Neuroimage Clin.* 19, 813–823. doi: 10.1016/j.nicl.2018.05.036

Kirschner, S. M., and Galperin, G. J. (2001). Psychiatric defenses in New York county: pleas and results. *J. Am. Acad. Psychiatry Law* 29, 194–201.

Koenig, E. G. (2016). A fair trial: when the constitution requires attorneys to investigate their clients' Brains. *Fordham Urban Law J.* 31, 177–225.

Kornhuber, H. H., and Deecke, L. (1965). Hirnpotentialänderungen bei willkürbewegungen und passiven bewegungen des menschen: bereitschaftspotential und reafferente potentiale. *Pflüger's Archiv Gesamte Physiologie Menschen Tiere* 284, 1–17. doi: 10.1007/bf00412364

Lerner, J. S., Li, Y., Valdesolo, P., and Kassam, K. S. (2015). Emotion and decision making. *Annu. Rev. Psychol.* 66, 799–823.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behav. Brain Sci.* 8, 529–539. doi: 10.1017/s0140525x00044903

Libet, B. (1999). Do we have free will? *J. Conscious. Stud.* 6, 47–57.

MacArthur (2019). *Research Network on Law and Neuroscience* [Online]. Available online at: http://www.lawneuro.org/ (accessed December 2, 2019).

Mason, D. A., and Frick, P. J. (1994). The heritability of antisocial behavior: a meta-analysis of twin and adoption studies. *J. Psychopathol. Behav. Assess.* 16, 301–323. doi: 10.1007/bf02239409

Mayberg, H. (2010). "Does neuroscience give us new insights into criminal responsibility," in *A Judge's Guide to Neuroscience: A concise Introduction*, ed. M. Gazzaniga (Santa Barbara, CA: University of California), 37–41.

McCabe, D. P., and Castel, A. D. (2008). Seeing is believing: the effect of brain images on judgments of scientific reasoning. *Cognition* 107, 343–352. doi: 10.1016/j.cognition.2007.07.017

McCabe, D. P., Castel, A. D., and Rhodes, M. G. (2011). The influence of fMRI lie detection evidence on juror decision−making. *Behav. Sci. Law* 29, 566–577. doi: 10.1002/bsl.993

Miller, G. (2010). Brain exam may have swayed jury in sentencing convicted murderer. *Science*. Available online at: https://www.sciencemag.org/news/2010/12/brain-exam-may-have-swayed-jury-sentencing-convicted-murderer (accessed November 15, 2019).

Morse, S. J. (2006). Brain overclaim syndrome and criminal responsibility: a diagnostic note. *Ohio State J. Crim. Law* 3, 397–412.

Morse, S. J. (2015). "Neuroscience, free will, and criminal responsibility," in *Faculty Scholarship Paper 1604*, ed. W. Glannon (Philadelphia, PA: University of Pennsylvania Law School: Penn Law: Legal Scholarship Repository).

Mulvaney, M. K., and Mebert, C. J. (2007). Parental corporal punishment predicts behavior problems in early childhood. *J. Fam. Psychol.* 21, 389–397. doi: 10.1037/0893-3200.21.3.389

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2005). Surveying freedom: folk intuitions about free will and moral responsibility. *Philos. Psychol.* 18, 561–584. doi: 10.1080/09515080500264180

Napier, S. (2019). The minimally conscious state, the disability bias, and the moral authority of advance directives. *Int. J. Law Psychiatry* 65:101333. doi: 10.1016/j.ijlp.2018.03.001

Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science* 331, 1401–1403. doi: 10.1126/science.1192931

NIDA (2019). *The Science of Drug Use and Addiction: The Basics* [Online]. North Bethesda, MD: The National Institute on Drug Abuse.

Pallarés-Dominguez, D., and Gonzalez Esteban, E. (2016). The ethical implications of considering neurolaw as a new power. *Ethics Behav.* 26, 252–266. doi: 10.1080/10508422.2015.1012763

Pardo, M. S., and Patterson, D. (2010). Philosophical foundations of law and neuroscience. *Univ. Ill. Law Rev.* 2010, 1211–1250.

Perlin, M. L. (2016). "The insanity defense: Nine myths that will not go away," in *The Insanity Defense: Multidisciplinary Views on its History, Trends and Controversies*, ed. M. D. White (New York, NY: New York Law School).

Pizzetti, F. G. (2011). In quest of constitutional principles of "Neurolaw". *Medicina Secoli* 23, 963–990.

Randa, L. E. (1997). *Society's Final Solution: A History and Discussion of the Death Penalty*. Lanham, MD: University Press of America.

Roskies, A. L., Schweitzer, N. J., and Saks, M. J. (2013). Neuroimages in court: less biasing than feared. *Trends Cogn. Sci.* 17, 99–101. doi: 10.1016/j.tics.2013.01.008

Sapolsky, R. M. (2017). *Behave: The Biology of Humans at Our Best and Worst*, Chapter 16. New York, NY: Penguin Random House, 580–613.

Satel, S., and Lilienfeld, S. O. (2014). Addiction and the brain-disease fallacy. *Front. Psychiatry* 4:141. doi: 10.3389/fpsyt.2013.00141

Schaler, J. A. (2000). *Addiction is a Choice*. Peru, IL: Carus by Open Court.

Schultze-Kraft, M., Birman, D., Rusconi, M., Allefeld, C., Görgen, K., Dähne, S., et al. (2016). The point of no return in vetoing self-initiated movements. *Proc. Natl. Acad. Sci. U.S.A.* 113, 1080–1085. doi: 10.1073/pnas.1513569112

Schweitzer, N. J., and Saks, M. J. (2011). Neuroimage evidence and the insanity defense. *Behav. Sci. Law* 29, 592–607. doi: 10.1002/bsl.995

Schweitzer, N. J., Saks, M. J., Murphy, E. R., Roskies, A. L., Sinnott-Armstrong, W., and Gaudet, L. M. (2011). Neuroimages as evidence in a mens rea defense: no impact. *Psychol. Public Policy Law* 17, 357–393. doi: 10.1037/a0023581

Shaw, E. (2018). Counterproductive criminal rehabilitation: dealing with the double-edged sword of moral bioenhancement via cognitive enhancement. *Int. J. Law Psychiatry* 65:101378. doi: 10.1016/j.ijlp.2018.07.006

Slobogin, C. (2011). Prevention as the primary goal of sentencing: the modern case for interdeterminate dispositions in criminal cases. *San Diego L. Rev.* 48:1127.

Slobogin, C. (2017). Neuroscience nuance: dissecting the relevance of neuroscience in adjudicating criminal culpability. *J. Law Biosci.* 4, 577–593. doi: 10.1093/jlb/lsx033

Soon, C. S., Brass, M., Heinze, H.-J., and Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nat. Neurosci.* 11, 543–545. doi: 10.1038/nn.2112

Spranger, T. E. (2012). *International Neurolaw: A Comparative Analysis*. Heidelberg: Springer-Verlag.

Steele, V. R., Claus, E. D., Aharoni, E., Vincent, G. M., Calhoun, V. D., and Kiehl, K. A. (2015). Multimodal imaging measures predict rearrest. *Front. Hum. Neurosci.* 9:425. doi: 10.3389/Fnhum.2015.00425

Unnever, J. D., Cullen, F. T., and Applegate, B. K. (2005). Turning the other cheek: reassessing the impact of religion on punitive ideology. *Justice Q.* 22, 304–339. doi: 10.1080/07418820500089091

Vergara, V. M., Mayer, A. R., Damaraju, E., Kiehl, K. A., and Calhoun, V. (2017). Detection of mild traumatic brain injury by machine learning classification using resting state functional network connectivity and fractional anisotropy. *J. Neurotrauma* 34, 1045–1053. doi: 10.1089/neu.2016.4526

Vincent, N. A. (2013). *Neuroscience and Legal Responsibility*. New York, NY: Oxford University Press.

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., and Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* 368, 1388–1397. doi: 10.1056/NEJMoa1204471

Walker, N. (1985). The insanity defense before 1800. *Ann. Am. Acad. Polit. Soc. Sci.* 477, 25–30. doi: 10.1177/0002716285477001003

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., and Gray, J. R. (2008). The seductive allure of neuroscience explanations. *J. Cogn. Neurosci.* 20, 470–477. doi: 10.1162/jocn.2008.20040

Weisberg, R. (2005). The death penalty meets social science: deterrence and jury behavior under new scrutiny. *Annu. Rev. Law Soc. Sci.* 1, 151–170. doi: 10.1146/annurev.lawsocsci.1.051804.082336

Yeung, W., Goto, T. K., and Leung, W. K. (2017). A bibliometric review of research trends in neuroimaging. *Curr. Sci.* 112:725. doi: 10.18520/cs/v112/i04/725-734

Yoshikawa, H., Aber, J. L., and Beardslee, W. R. (2012). The effects of poverty on the mental, emotional, and behavioral health of children and youth: implications for prevention. *Am. Psychol.* 67, 272–284. doi: 10.1037/a0028015

Zeki, S., Goodenough, O., and Sapolsky, R. M. (2004). The frontal cortex and the criminal justice system. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 359, 1787–1796.

Zimmer, C. (2005). *Soul Made Flesh: The Discovery of the Brain–and How it Changed the World*. New York, NY: Simon and Schuster.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

REVIEW
published: 16 April 2020
doi: 10.3389/fpsyg.2020.00695

The Empathic Brain of Psychopaths: From Social Science to Neuroscience in Empathy

*Josanne D. M. van Dongen**

Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, Netherlands

Empathy is a crucial human ability, because of its importance to prosocial behavior, and for moral development. A deficit in empathic abilities, especially affective empathy, is thought to play an important role in psychopathic personality. Empathic abilities have traditionally been studied within the social and behavioral sciences using behavioral methods, but recent work in neuroscience has begun to elucidate the neural underpinnings of empathic processing in relation to psychopathy. In this review, current knowledge in the social neuroscience of empathy is discussed and a comprehensive view of the neuronal mechanisms that underlie empathy in psychopathic personality is provided. Furthermore, it will be argued that using classification based on overt behavior, we risk failing to identify important mechanisms involved in the psychopathology of psychopathy. In the last decade, there is a growing attention in combining knowledge from (neuro)biological research areas with psychology and psychiatry, to form a new basis for categorizing individuals. Recently, a converging framework has been put forward that applies such approach to antisocial individuals, including psychopathy. In this bio-cognitive approach, it is suggested to use information from different levels, to form latent categories on which individuals are grouped, that may better reflect underlying (neurobiological) dysfunctions. Subsequently, these newly defined latent categories may be more effective in guiding interventions and treatment. In conclusion, in my view, the future understanding of the social brain of psychopaths lies in studying the complex networks in the brain in combination with the use of other levels of information (e.g., genetics and cognition). Based on that, profiles of individuals can be formed that can be used to guide neurophysiological informed personalized treatment interventions that ultimately reduce violent transgressions in individuals with psychopathic traits.

Keywords: psychopathy, empathy, theory of mind, social neuroscience, complex brain networks, forensic mental health

OPEN ACCESS

Edited by:
Eric García-López,
Instituto Nacional de Ciencias Penales, Mexico

Reviewed by:
Yu Gao,
Brooklyn College (CUNY), United States
Fernando Barbosa,
University of Porto, Portugal

***Correspondence:**
Josanne D. M. van Dongen
j.d.m.vandongen@essb.eur.nl

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *28 November 2019*
Accepted: *23 March 2020*
Published: *16 April 2020*

Citation:
van Dongen JDM (2020) The Empathic Brain of Psychopaths: From Social Science to Neuroscience in Empathy. Front. Psychol. 11:695. doi: 10.3389/fpsyg.2020.00695

INTRODUCTION

Empathy is seen as the "natural capacity to share, understand, and respond with care to the affective states of others" (Decety, 2012). It plays an important role in social interactions, not only in humans, but also other species including apes (de Waal, 2012), and rodents (Decety et al., 2016). Moreover, empathy is thought to play an important role in affecting prosocial behavior, inhibiting aggressive

behavior and is found to be fundamental to the development of moral behavior (Eisenberg and Eggum, 2009). Over centuries of literature on empathy has shown that empathy is sometimes confused with, or used interchangeably with other concepts, such as sympathy and compassion. In my view, empathy encompasses different facets and differs from sympathy and compassion in that empathy not only includes *other-oriented* empathy (i.e., empathic concern), but also entails *self-oriented* responses (i.e., emotional distress and emotional contagion). Thus, empathy differs from sympathy and compassion in the sense that it includes feelings that are similar *as* the other feels and not feelings *for* how the other person feels (Batson, 2009).

Since social sciences are concerned with different disciplines that examine society and how individuals interact with the social environment, empathy was originally studied within these disciplines. Psychology, the study of the human behavior and mind, has naturally focused on behavioral aspects of social interactions. For instance, behavioral research in social psychology has led to the empathy-altruism hypothesis (Batson, 2009). This hypothesis is supported on ample evidence that empathy is an other-oriented behavior, and is not egoistic in nature. Moreover, it is suggested that empathic concern for others results in altruistic motivation to care and help others.

Importantly, empathy is such an essential component of healthy human social interactions that absence of it may lead to severe social and cognitive dysfunctions. A personality structure often marked by a lack of empathy is psychopathic personality. Thus, clinical psychology is also concerned with the process of empathy and how this ability influences antisocial personality (including psychopathy) and behavior. And although manifestations of personality and psychopathology in psychology is originally studied from a behavioral point of view (i.e., symptoms), psychological science is integrating the neurobiological underpinnings of cognition and behavior. Also, in the study of psychopathic personality, scholars become more aware of the fact that psychopathic personality is heterogeneous, consisting of multiple facets of traits with each of these traits having different underlying neuro-cognitive processes.

Alternative approaches to study personality and psychopathology have emerged decades ago (see for example Morton and Frith, 1995 on causal modeling). Also, approaches that incorporate neuroscience, such as the Research Domain Criteria (RDoC; Insel et al., 2010), have emerged already in the nineteenth century (Arzy and Danziger, 2014). However, these approaches have not been applied to the study of psychopathic personality more specifically, only until a couple of years ago (Blair, 2015a,b; Brazil et al., 2018). The idea behind these approaches is that mental disorders are originally classified based on behavioral symptoms (e.g., DSM criteria), but that, during the last decades, it has become increasingly apparent that these disorders consist of dysfunctional bio-cognitive processes related to different processes at the neural level. Each of these processes are found to be existent transdiagnostically, and therefore must be studied individually. In the case of empathy, this is not only dysfunctional in psychopathic personality, but also in autism, schizophrenia, and borderline personality disorder (Lockwood, 2016).

Thus, elucidation of the neural underpinnings of empathy will not only help us understand social interactions, but also help us understand the neural and cognitive mechanisms of emotion processing, motivation (i.e., empathic concern), and individual differences in antisocial and psychopathic personality.

The aim of this review paper is to give an overview of our current knowledge on the role of neuroscience in the study of empathy in psychopathic personality. First, some conceptual matters of empathy and associated concepts are clarified, and it is argued that the construct of empathy needs to be defined by several subcomponents and processes that are underpinned by diverse processes in the brain. Additionally, studies on the neural circuits involved in empathy are reviewed. Next, a short historical overview of psychopathy as a construct will be given, as well as different theoretical models on this personality. In the third section, a review of empirical evidence is given that supports the role of social neuroscience in psychopathic personality. Finally, I will discuss a new way forward in using neuroscience in the study of the "empathic brain" of psychopaths.

EMPATHY

As already mentioned in the introduction, the term "empathy" is applied to various phenomena, including feeling the same as another person is feeling, feeling pity for another person, and knowing what the other person is feeling or thinking (Batson, 2009). The labels of these concepts also vary between empathy, sympathy, pity, and compassion. Although these concepts are related, and sometimes overlap, they do not represent the same psychological (and neurobiological) phenomena. Not surprisingly, there is still a debate on what the construct of "empathy" entails. Some scholars include both self- and other-oriented processes (Decety, 2010), and others only include those phenomena that are oriented toward the person in need (other-oriented; empathic concern; Batson, 2009).

Hence, as already briefly outlined above, empathy (the capacity to understand and know the difference between one's own emotions and feelings and that of another person) is distinguished from sympathy (to be concerned about the wellbeing of another person). While the terms empathy and sympathy are often used interchangeably, the two can be differentiated: the experience of empathy can lead to different outcomes: an *other-oriented* motivation, sympathy, or a *self-oriented* feeling of distress imposed by the stressor which includes, and may also be congruent to the emotional state of that other person (emotional contagion). Sympathy may be the result of understanding another's affective state but does not have to be consistent with that state. Given the complexity of the experience of empathy, it is important to first break down this construct into component processes.

The Components of Empathy

Generally, researchers have postulated that empathy includes both affective and cognitive components (Decety and Jackson, 2004; Eisenberg and Eggum, 2009; Decety, 2010; Zaki and Ochsner, 2012). Based on evidence from cognitive neuroscience

and developmental psychology, a number of different, but interacting mechanisms result in the experience of empathy (Decety, 2010; Zaki and Ochsner, 2012): (1) An affective component of affective sharing or emotional contagion; a bottom-up process which is a result of perception-action coupling, and emotion perception (Preston and de Waal, 2002). (2) A cognitive aspect of mentalizing or perspective taking (i.e., Theory of Mind; ToM); the ability to make a distinction between oneself- and other, and (3) executive functions which influence the extent of an empathic experience, and results in empathic concern (i.e., sympathy), using amongst others the perceiver's motivation, memories, and intentions.

Research indicates that the affective empathy develops before cognitive empathy. Following the Perception-Action Model (Preston and de Waal, 2002), it is suggested that newborns are able to mimic facial expressions, and infants are found to become distressed if they hear another baby cry. That is, they perceive the crying of another infant that (automatically) contributes to affective sharing. Thus, affective responsiveness is present at an early age, is automatic, and is the result of mimicry and somato-sensorimotor resonance between the self and other.

The cognitive components of empathy include ToM, or mentalizing. This is the ability to infer the mental states of another person, which includes executive functions such as attention, working memory, and self-regulation. These "higher" cognitive abilities are suggested to develop later in life, because the prefrontal cortex develops more slowly than more basal (emotion related) brain areas, reaching maturation in late adolescence (Bunge et al., 2002). The development of the prefrontal cortex permits children to express their feelings and develop self-regulation by using inhibitory control over their thoughts, attention, and actions (Diamond, 2002). Thus, although affective aspects of empathy develop early in life, maturation of the frontal brain influences the way executive functions interact with empathic responding. That is, executive functions (i.e., emotion regulation, inhibitory control, etc.) have their effect on how empathy develops in its full scope of facets.

Although at first it was thought that ToM abilities develop later in childhood, more recent studies have suggested that babies already have obtained these abilities to some extent by the age of 4 years (Onishi and Baillargeon, 2005). Moreover, babies as young as 7 months are found to have a "social sense" (Kovács et al., 2010). This social sense is an automatically computed *online* belief about another agent, which is maintained even in the absence of that agent.

Sharing Emotions With Others

The perception and resonance of the affective states of another person are thought to result in shared representations of oneself and others. Evidence suggests that for particular emotions, such as fear, disgust and pain, there are brain regions that map the emotions of another to oneself. That is, we not only "simply" understand the emotions of another person, we also *feel as* and *feel with* the other person. These abilities are found to be grounded in shared representations (Keysers and Gazzola, 2006). However, although the human mind has, in some cases, an egocentric bias (we think that others think and feel as we think and feel), successful social interactions partly result from the ability to distinguish oneself from the other (Sommerville and Decety, 2006).

The shared-representation theory of social cognition (Sommerville and Decety, 2006) suggests that the experience of emotion in oneself and the perception of another's emotions draw on many of the same underlying neural circuits and computational processes, including somatosensory and motor representation (see later in this review for neural structures and mechanisms involved in empathy). As will be discussed later, one important mechanism involved in this shared representation, is the mirror neuron system (Rizzolatti and Craighero, 2004; Iacoboni et al., 2005).

Past research generally has focused on "what is shared" by these shared representations (i.e., cognition and/or emotional states), and less on "how these are shared." Advances have been made by Bird and Viding (2014), who formulated a model of mechanisms by which the affective state in another may result in an empathic response in the self. In this Self Other Model of Empathy (SOME), empathy is differentiated from emotional contagion in that emotional contagion results from the vicarious experience of the affective state of another person, *without* recognizing this state as being a part from that other person. Empathy results from the mechanisms of emotional contagion, with the addition that one recognizes that the experienced affective state is experience by that other person. This accomplished by a so-called Self/Other switch, a system that requires information from the ToM system to results in a switch from self (the default) to the other (Bird and Viding, 2014). Together with understanding the situation both the self and the other are in, it evaluates whether the affective state of the self, corresponds to the situation and emotional state of the other person.

Neural Circuits in Empathy

Neuroscientists have started to elucidate the neurobiological underpinnings of empathy (Decety, 2010; Zaki and Ochsner, 2012). Functional neuroimaging studies have shown that imagining emotional experiences from our own and from someone else's perspective result in comparable psychophysiological reactions and patterns of brain activation. For example, Ruby and Decety (2004) presented participants with short written scenario's depicting real-life situations (e.g., someone opens the toilet door that you have forgotten to lock) which induce social emotions (e.g., shame, guilt, pride), as well as emotionally neutral situations. Subsequently, they asked them to imagine how they would feel if they were in those situations, and how their mother would feel in those situations. Results showed that the imagined emotional conditions for both the self and the other perspectives led to similar activation of brain areas that are involved in emotional processing, including the amygdala and the temporal poles.

In a study by Preston et al. (2007), heart rate, skin conductance, and neuroimaging measurements were combined in participants who were also asked to imagine a personal experience of fear or anger from their own past, and an equivalent experience from another person as if it were happening to

them. Results confirmed earlier results, in that similar patterns of psychophysiological and neurological activation were found when participants could relate to the scenario of the other, and to those of personal emotional imagery.

Developmentally, the process of empathic distress or emotional distress may play a role in the underpinnings of prosocial behavior (Hoffman, 1990). Also, the expression of pain offers an important signal to others, that motivates behavior such as caring for a person in distress (i.e., sympathy). It is the affective experience of pain that indicates an aversive state and motivates behavior that, for example ends, or reduces exposure to the source that has led to the aversive state in the first place (Price, 2000). The perception and experience of pain is therefore often used by researchers as a valuable and ecologically valid means to investigate the experience of empathy.

Following the above, most research in empathy has focused on empathy for pain, and how different factors modulate its experience and behavioral expressions (Singer and Lamm, 2009; Lamm et al., 2011). For instance, as was already indicated in the paragraph above, different functional neuroimaging studies have shown that similar brain regions are activated during the personal experience of pain and when attending to the pain of others (Lamm and Majdandžiæ, 2015; Zaki et al., 2016). These regions include the anterior insula (AIC), anterior mid and dorsal anterior cingulate cortex (ACC), and periaqueductal gray (Lamm et al., 2011). In one functional magnetic resonance imagining (fMRI) experiment, participants were scanned during a condition of feeling a moderately painful pinprick stimulus to the fingertips and another condition in which they watched another person's hand undergo similar stimulation (Morrison et al., 2004). Both conditions resulted in increased activity in the right dorsal ACC. Another fMRI study with healthy participants showed that the dorsal ACC, the AIC, cerebellum, and brain stem were activated both when the participants experienced a painful stimulus, as well as when they observed the same in another person receiving it. However, only the actual experience of pain resulted in activation in the somatosensory cortex and a more ventral region of the ACC (Singer et al., 2004). Additionally, these results are supported by two other fMRI studies (Jackson et al., 2005, 2006).

In a study by Zaki et al. (2007), participants were scanned while they received hurtful thermal stimulation (self-pain condition) or watched short videos of other people receiving painful stimulation (other pain condition). With connectivity analyses, the researchers found areas whose activity covaried with ACC and AI activity during self or other pain either across time (intra-individual connectivity) or across participants (inter-individual connectivity). Both connectivity analyses revealed clusters in the midbrain and periaqueductal gray with greater connectivity to the AI during self-pain as compared to other pain. Greater connectivity to the ACC and AI during other pain than during self-pain was found in the dorsal medial prefrontal cortex, using both types of analysis. Intra-individual connectivity analyses also revealed regions in the superior temporal sulcus, posterior cingulate, and precuneus that became more connected to ACC during other pain compared with self-pain. These and other results show that there are distinct neural networks associated with ACC and AI in response to personal experience of pain and response to seeing other people in pain (Morrison and Downing, 2007; Zaki et al., 2007).

Facial expressions of pain form an important category of facial expression that is easily comprehended by observers. In one study Botvinick et al. (2005), the neural response to these facial pain expressions were examined using fMRI while subjects viewed short video sequences showing faces expressing either moderate pain or, for comparison, no pain. Facial expressions of pain were found to lead to cortical activation similar to areas activated in firsthand experience of pain, including the ACC and AI. Similar results were found by Lamm et al. (2007), who scanned participants, and let them listen to painful sounds and let them watch videos of people expressing pain due to listening to painful sounds.

Concerning the brain structures involved in empathic experiences, the mirror-neuron system (MNS) and somatosensory cortex are suggested to be involved in experiencing and seeing the actual cause of pain (Decety, 2010). However, it remains debated whether the emotion sharing mechanism in humans actually requires the involvement of the MNS (Baird et al., 2011). Mirror neurons are a class of cells that were first identified in monkeys (Gallese et al., 1996). Although first it was thought that these cells were mainly involved in action understanding and imitation, now, different higher cognitive functions have been found to be associated to the MNS, including empathy (Rizzolatti and Craighero, 2004).

On the contrary, however, a conceptual analysis by Jacob (2008) of empirical research on mirror neurons and their assumed contribution to empathy, concluded that motor resonance (as a result of MNS activity), is neither necessary nor sufficient for representing another individual's intentions. It was argued that mirror neurons may be best interpreted as motor system facilitators (Hickok, 2009). Their involvement in empathy may then be via the so-called "mimicry" (Decety, 2010) that is suggested to be necessary for perception-action coupling (Preston and de Waal, 2002). Subsequently, the ACC and AIC are associated with the affective value of somatosensory stimuli within this emotion sharing network (Singer et al., 2004; Keysers et al., 2010).

In sum, previous functional neuroimaging studies indicate that perceiving or imagining another individual in pain is associated with activity in brain areas processing sensory, and motivational-affective dimensions of pain in oneself.

PSYCHOPATHY: AN OVERVIEW

Psychopathy is a personality consisting of characteristics including callousness, lack of guilt, shallow affect, impulsive and antisocial behavior (Cleckley, 1976). Approximately 1% of the general population, but 20–30% of the prison population are found to have a psychopathic personality (Hare, 1999). Because of their behavioral characteristics, psychopathic individuals pose great costs to society (i.e., economically, mental healthcare, and criminal justice), estimated at $400 billion in the USA alone (Kiehl and Buckholtz, 2010). This seems to be comparable within

European countries, such as the Netherlands, where treatment costs of antisocial offenders in forensic psychiatric facilities is $160,000 a year per person. These costs are extremely high, especially when compared to costs related to other diseases, such as treating type 2 diabetes, which is estimated at only $1,700–2,100 a year per person (Brandle et al., 2003).

Because of the high costs, both financially, but also emotionally, that psychopathic individuals pose, there is a strong need for classifying these individuals and developing treatment interventions that will target this personality. Unfortunately, as reflected by their high risk of recidivism, psychopathic individuals account for the majority of failed treatment efforts. Several attempts have been made to treat antisocial individuals, including those with psychopathic personality, using a variety of clinical approaches (Harris and Rice, 2006; Gibbon et al., 2010; Salekin et al., 2010). While there is some support for successfully targeting some characteristics of this personality using psychological and pharmacological treatment, there is no evidence that current treatments effectively address this personality. Therefore, some clinicians and researchers have postulated that individuals with elevated levels of psychopathy, maybe even untreatable (Harris and Rice, 2006). However, I think that the development of effective treatment interventions may be advanced by recognizing the heterogeneity of psychopathic personality and incorporating knowledge about the underlying neurobiological correlates of this personality into the development of more specific treatments.

Subtypes of Psychopathy

Cleckley's (1976) *The Mask of Sanity* served as a groundwork for different conceptualizations and measurements of psychopathic personality. Hare (1991, 2003) used Cleckley's description of clinical criteria as a basis for the development of a diagnostic instrument for the assessment of psychopathic personality. The Revised version of Hare's Psychopathy Checklist (PCL-R), an interview and file-based assessment instrument, is still regarded as the "golden standard" for assessing psychopathy in forensic and correctional settings. Generally, a score of 30 or above out of 40 (maximum score), is regarded as a cutoff for the classification as a psychopath. In European countries however, a cutoff score of 25 is being used. The PCL-R measures psychopathy in terms of two broad factors: Factor 1, including Affective and Interpersonal facets (i.e., grandiosity, deceitfulness, lack of empathy, and lack of remorse) of psychopathy, and Factor 2, including Antisocial and Lifestyle facets (i.e., deficit in behavioral inhibition and control).

Throughout the years, a lot of research has been conducted on the usefulness of the PCL-R and its different variants (Neumann et al., 2007). Like any assessment instrument, it has certain limitations. One is that several of its items refer directly to criminal activity, which makes the PCL-R less appropriate for use in non-correctional samples. Another is that the PCL-R is very time consuming to administer, and impractical for large scale data collection efforts because of its interview-based procedure and requirement of collateral (i.e., archival file) information. As a result, different other (self-report) measures are developed for the assessment of psychopathic personality during the years, some of them found to be more promising than others.

The term psychopathy has commonly been used as a unitary construct and to refer to one particular group of individuals scoring higher than a cut-off score on the PCL-R (Hare, 2003). The problem with assuming psychopathy as a unitary personality construct, is that it does not consider that persons scoring high and low on particular characteristics of psychopathy such as impulsivity, empathy and even anxiety are different from one another (Skeem et al., 2003). Nowadays, many researchers view psychopathic personality as being multidimensional, and believe that this personality includes multiple subtypes that differ significantly in etiology and personality characteristics (e.g., Skeem et al., 2003; Patrick et al., 2009).

During the last decades, different self-report measures of psychopathy are developed, to overcome some of the (practical) difficulties that come with the use of the PCL-R. These include for example the Self-Report Psychopathy Scale and its Short Form (SRP; Hare, 1980; SRP-SF; Paulhus et al., 2016), the Psychopathic Personality Inventory and its revised version (i.e., PPI; Lilienfeld and Andrews, 1996; and PPI-R; Lilienfeld and Widows, 2005), and the Levenson Self-Report Psychopathy Scale (LSRP; Levenson et al., 1995). One of the alternative frameworks of psychopathy that addresses the above multiple psychopathy types principle, is the Triarchic Psychopathy Model. Patrick et al. (2009) have proposed this conceptualization based on the observation that previous literature reveals three important facets within the construct of psychopathy: *boldness* (reduced emotionality, resilience to stress, and social dominance), *meanness* (lack of empathy, cruelty, and aggressive behavior toward others), and *disinhibition* (impulsivity and dysregulation of negative affect) (but see Roy et al., 2020 for a septarchic structure of this model). These three constructs are viewed as connected, yet distinct from one another, and can be measured and understood separately. The assumption is that the three dimensions can be combined to create descriptions for different subtypes of psychopathic personality. This approach also claims to account for adaptive features seen in psychopathy (i.e., boldness), traits that were incorporated in classic accounts of psychopathy (Cleckley, 1976; Lykken, 1995), which are not incorporated in the PCL–R. The construct of meanness, but also boldness to some extent, has theoretical relations with the concept of empathy. While meanness is viewed as the core construct associated with a lack of affective empathy (Sellbom and Phillips, 2013; Stanley et al., 2013), the concept of boldness does also entail fearlessness and the ability to remain calm in the face of threat, suggesting a negative relation to the personal distress facet of empathy. However, for an individual to show these boldness traits, this individual also needs to have (high) functioning mentalizing ability to successfully manipulate others.

SOCIAL NEUROSCIENCE OF EMPATHY IN PSYCHOPATHY

Theoretical Accounts

As described in previous paragraphs, individuals scoring high on psychopathic traits are defined as fearless, callous and have a lack of empathic disregard for others combined with impulsive and

antisocial behavior (Hare, 2003). Also, it is found that they have difficulty controlling their emotions and often lack fear when facing punishment. Insights into neural circuits underpinning healthy empathic behavioral processes may shed light on potential neural dysfunctions in psychopathic personality. Conversely, advances made in the description of the component processes underlying psychopathic personality are invaluable as a complement to other methods of empathy research.

Different accounts have been formulated that explain psychopathic personality and its consecutive behavior. On the one hand are accounts that explain psychopathic personality on the basis of deficits in emotions, most notably anxiety and fear. In these theories it is argued that psychopathic individuals lack fear responses when faced with stressful situations and therefore do not form punishment related associations (Fowles, 1988; Patrick et al., 1994; Lykken, 1995). These theories are based on research that has shown deficits in emotion recognition (Marsh and Blair, 2008; Dawel et al., 2012), and (neuro)physiological responses to fear (Patrick et al., 1994; Kiehl et al., 2001).

On the other hand are accounts that are based on attentional deficits (i.e., the Response Modulation Hypothesis; Newman et al., 1987; Newman, 1998). In these theories, it is argued that deficits in psychopathic personality relate to difficulties in reallocating attention to information that is not relevant when engaged in goal-directed behavior. These attention views are partly based on findings that have shown that fear deficits seen in psychopathy are moderated by attention (Newman et al., 2010; Baskin-Sommers et al., 2011).

The Integrated Emotion Systems (IES) model (Blair, 2007, 2013), follows work that has been done within the emotion deficits approach, such as work from Patrick et al. (1994). This model stresses the importance of the amygdala. Research has shown that the amygdala is critical for stimulus-reinforcement learning, for example in aversive conditioning, which is impaired in psychopathy (Rothemund et al., 2012). This finding corresponds to findings that have shown that psychopaths show reduced activation of the amygdala during aversive conditioning (i.e., Birbaumer et al., 2005). In addition, the IES model also stresses the importance of the ventro-medial prefrontal cortex (vmPFC) including the orbitofrontal cortex (OFC).

Following this, according to the IES model, processing of emotional stimuli is involved in (moral) behavioral transgressions. Transgressions are learned to be considered as "bad" because of the aversive feedback that follows that transgression, for example the distress of the victims of these transgressions. Impaired stimulus-reinforcement learning as the result from amygdala dysfunction, and impaired responsiveness to the distress of others (e.g., communicated by facial expressions; Blair, 2011) lead to deficits in empathy for others and subsequently to (moral) behavioral transgression.

In support of the IES model, the amygdala is found to be important for processing expressions of fear and distress (Murphy et al., 2003), and individuals with psychopathy who are violent show reduced amygdala responses to fearful expressions (Dolan and Fullam, 2009). This dysfunctional response reflects a dysfunction in empathic responding (i.e., personal distress). Consequently, dysfunction in stimulus-reinforcement learning, thus learning the consequences (fear expression) of one's actions (aggression), results in a deficient response to transgressions (i.e., empathic concern). Different studies found reduced amygdala responses follow moral transgressions and moral decision-making in individuals with psychopathic traits (Glenn et al., 2009; Harenski et al., 2010).

In line with the IES model, the violence inhibition model (VIM; Blair, 1995, 2001) also views empathy as an important mechanism for moral socialization. The VIM in addition accounts for the inhibition of violent behavior (or the lack of inhibition of that behavior) by coupling the activation of the mechanism by distress cues with representations of the acts which caused the distress cues (i.e., transgressions). A child that is developing appropriately thus initially finds the pain of others' aversive and then, through aversive conditioning (or stimulus reinforcement), transgressions are inhibited because of the aversive consequences of that action. According to the VIM, individuals with psychopathic personality have dysfunctional neural circuits (i.e., the amygdala and vmPFC) involved in these associative learning mechanisms (Blair, 2001).

In support of the above, Greene et al. (2001) found that personal as opposed to impersonal moral choices let to increased vmPFC activity. Likewise, Luo et al. (2006) showed that in response to more severe moral transgressions, amygdala and vmPFC activity was increased when compared to less severe moral transgressions.

Following the IES model and the VIM, Blair (2007, 2008) argues that, while the amygdala is particularly involved with emotional responding and forming the learning basis of necessary for caring for the welfare of others, the vmPFC is particularly involved with the decision process following input from the amygdala. This corroborates with the idea that affective empathy (i.e., affective arousal/personal distress) is found to be mediated by subcortical structures from the limbic system, such as the amygdala. And, emotional decision-making, and subsequently empathic concern for others (including moral cognitions), are found to be mediated by the vmPFC (Decety, 2010).

Functional Neuroimaging Studies

Neuroimaging studies found that above mentioned structures relevant for empathy are dysfunctional in persons with psychopathic traits (e.g., Koenigs et al., 2007; Shamay-Tsoory et al., 2010; Decety et al., 2013b; and see Lockwood, 2016 for a review). For instance, in one study, persons scoring high and low on the PCL-R were examined during the viewing of pictured depicting bodily harm (Decety et al., 2013a). They had to imagine that this harm involved oneself, or another person. During the imagine-self perspective, participants with higher scores on psychopathy showed atypical response in the AI, aMCC, SMA, IFG, somatosensory cortex, and right amygdala. This corresponds with the brain network involved in the experiencing of pain. Conversely, during the imagine-other perspective, individuals with higher scores on psychopathy showed a different pattern of cortical activation and effective connectivity resulting from the AI and amygdala with the OFC and vmPFC. Moreover, the imaging-other condition, response

in the amygdala and insula was inversely correlated with the interpersonal and affective traits of psychopathy.

Meffert et al. (2013) conducted a study using fMRI involving the viewing of scenarios depicting hand movements and found a similar pattern of reduced activation of brain areas involved in empathy in persons with psychopathy compared with controls. Interestingly however, they also found that when these individuals were instructed to empathize with the person in the videos, the reduction in activation became less. The authors concluded that persons with psychopathy do not have a total absence or incapacity to empathize with another person, but that brain mechanisms involved are not automatically activated in these individuals (see also Keysers and Gazzola, 2014 on the ability vs. propensity for empathy). That persons with psychopathic traits do not seem to have a total lack of empathy was also shown by a recent online survey study (Kajonius and Björkman, 2020). In this study, the authors investigated the disposition of empathy and the ability to empathize in persons scoring higher and lower on the Dark Triad personalities (i.e., Machiavellianism, psychopathy, and narcissism). It was found that dark triad personality was not related to ability-based empathy, but strongly negatively related to dispositional based empathy.

With respect to the different facets that make up empathy and psychopathy, it may be of importance that most research that support a lack of empathy in psychopathy are supporting a lack of affective empathy. Robinson and Rogers (2015) for example, found that psychopathic criminals had no impairment in cognitive empathy (i.e., ToM or mentalizing), but did not seem to possess affective empathy. Likewise, Sandoval et al. (2000) found a negative relationship between self-reported affective empathy and psychopathy, but no relationship with cognitive empathy. However, there are also studies in which no relations or negative associations were found between both affective and cognitive empathy and psychopathy (Brook et al., 2013; Brook and Kosson, 2013; Domes et al., 2013).

Though ToM has been regarded as a cognitive aspect of empathy, according to the theoretical framework of Shamay-Tsoory et al. (2010), ToM is a construct that can be separated into cognitive and affective aspects. Cognitive ToM resembles what is generally referred to as metalizing, while the affective part refers to the ability to infer on other's feelings and therefore relates to both affective and cognitive empathy. It is important to note that affective ToM differs from affective empathy, in that affective empathy also includes emotional contagion (feeling the same feeling as the other person does), while affective ToM does not.

Thus, when interpreting previous findings concerning the relation between psychopathy and empathy (including ToM), it is important to recognize the above mentioned difference in cognitive and affective ToM. As previously stated, most research found no lack of cognitive empathy in psychopathic individuals (Blair, 1996; Richell et al., 2003; Dolan and Fullam, 2009), while Brook and Kosson (2013) did find a lack of ToM in psychopaths. However, this lack of ToM concerned only negative emotions such as fear and sadness, which now would be interpreted as a lack of affective ToM, and not a deficit in cognitive ToM. Dysfunctions in ToM in persons with psychopathic traits are thus subtle and may be interpreted in a way that is not done so in previous studies.

The Default Mode Network

Throughout the years, studies examining neuronal networks involved in psychopathic personality have increasingly been carried out, for example by using functional connectivity analysis. Functional connectivity is defined as the relation between the neuronal activation patterns of anatomically separated brain areas. Psychopathy has mostly been associated with atypical functional connectivity in (regions of) the default mode network (DMN; Raichle, 2015), including the mPFC, posterior cingulate cortex, precuneus, and angular gyrus, as well as bilateral IPL expanding to posterior temporal areas around the TPJ (Buckner et al., 2008). The DMN has been implicated in empathy, self-processing and moral behavior (Buckner et al., 2008; Andrews-Hanna et al., 2010; Li et al., 2014), and abnormal functioning of this network may play an important role in explaining core psychopathic traits, such as impaired emotion recognition (e.g., affective ToM; Grimm et al., 2009), and impaired moral decision making (Tassy et al., 2013). Subsequently, the DMN now is becoming increasingly recognized as a network of the social brain (Mars et al., 2012).

To sum up, given the above reviewed literature, we may conclude that individuals with psychopathic traits are found to have a deficit in dispositional empathy, particularly related to the processing of distress and negative arousal cues (i.e., affective empathy and affective ToM). These deficits are likely to be related to dysfunctions in a wide brain network involved in empathy, including the vmPFC/OFC and amygdala. And because a lack of sharing of vicarious negative arousal in these individuals, this may result in not showing empathic concern for others. In other words, individuals with higher levels of psychopathic traits show weaker psychophysiological reactions to these negative arousal cues and have poor aversive conditioning and stimulus-reinforcement learning. However, it is important to mention some limitations to the above conclusion. One is that other brain systems are also important in mediating other psychopathic personality traits, such as impulsivity and other impairments in executive functioning (see Koenigs et al., 2011 for a review). However, reviewing these traits is not within the scope of this review on the social brain.

Also important, studies reviewed in this review largely involved neuroimaging studies using fMRI. Within the social neuroscience of empathy in psychopathic personality, studies using electrophysiological measurements are scarcer. Electrophysiological studies are of additional value here, for example because it gives insight in the functional dynamics of different processes in higher temporal resolution compared to fMRI. Also, studies involving empathy mainly have focused on empathy for pain. For future research, it is very important to elucidate further the electrophysiological correlates of empathy in relation to psychopathic traits using ecologically more valid stimuli in tasks, such as pictures depicting aggressive situations (see for example van Dongen et al., 2018), but also other forms of empathy, for instance "positive empathy" (see Morelli et al., 2015). When doing so, this gives more insight in the social

neuroscience aspects of empathy, not only the sensory aspects when the processing of pain stimuli is involved. Moreover, using aggression scenes or pictures depicting victims in distress is of particular importance, because of its ecological value when studying psychopathic personality.

THE MISSING LINK: THE WAY FORWARD

Research has mainly relied on social- and behavioral sciences when studying psychopathic personality. This makes sense, because psychopathic personality manifests itself most apparently at the surface with behavior that deviates from the social norm. Also, as with some forms of psychopathology, psychopathic personality has been generally viewed as a *mental* disorder. Though, as became clear in the current review, a shift from investigating forensic and correctional samples to community-based samples, accompanied by a shift from a diagnostic to dimensional perspective of psychopathic traits, has long been underway. Also, using classification based on overt behavior, we risk failing to identify important mechanisms involved in the psychopathology of psychopathic personality traits. For instance, assessments and tasks that are used to assess levels of empathy in this personality may not be sensitive enough to detect particular deficits in empathic abilities (Shamay-Tsoory et al., 2010; Domes et al., 2013). Thus, although the general view is that psychopaths lack affective empathy and have intact ToM, this may be challenged when using more sensitive ToM tasks. Moreover, when no overt behavioral differences between individuals scoring high and low on psychopathic traits are found, this may not automatically reflect "true" underlying resemblance in neurophysiological mechanisms. Also, when no behavioral differences are found, but underlying automatic (neural) processes differ in individuals with psychopathic traits, this may affect automatic responding outside the laboratory (e.g., Meffert et al., 2013). This points to the idea that, when necessary, psychopaths may use covert (computational) strategies in the brain to overcome otherwise automatic inappropriate responding.

In addition, as in this review discussed, complex and multifaceted nature of psychopathic personality, it is crucial to use additional neuroscientific insights to understand an individual (assessment) and for subsequent (targeted) effective treatment of higher levels of psychopathic personality. It has become clear that without neuroscience, the possibility to form a complete picture of psychopathologies and personalities, including psychopathic personality, is clearly missed. Hence, like mental disorders (Insel and Cuthbert, 2015), psychopathy now can be viewed as a disorder of the brain. Also, the influence of neuroscience in social science is not only important for a better understanding of the etiology, different expressions, and phenotypes of psychopathy, but also for the development of effective interventions. Because of the trial and error nature of interventions to date, much of these interventions are found not to be much effective (e.g., Salekin et al., 2010). By elucidating the underlying mechanisms that motivate persons with psychopathic traits in their behavior, interventions can be developed more targeted at specific dysfunctional mechanism, such as deficient dispositional empathy.

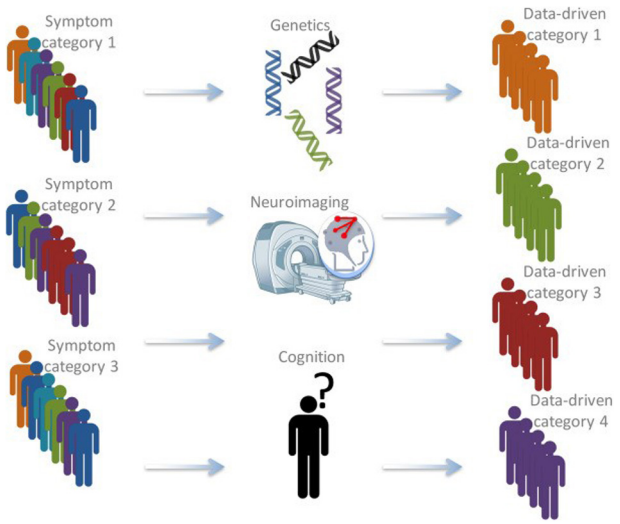


FIGURE 1 | Schematic depiction of the bio-cognitive approach (after Insel and Cuthbert, 2015). Currently, patients are often categorized based on behavioral symptoms. Using information from a variety of approaches, including genetics, structural and functional neuroimaging, cognitive measures, and computational psychiatry, latent categories that might be much better at grouping different patients and predicting therapy outcomes might be found (Brazil et al., 2018).

During the last decade, insights from (neuro)biology with psychology and psychiatry are increasingly combined to form a new basis for categorizing individuals (see **Figure 1**). Most prominent is the approach that has been put forward in the Research Domain Criteria (RDoC) framework, developed by the National Institute of Mental Health (NIMH; Insel et al., 2010). This framework aims to understand mental illness as the interaction of factors at multiple levels (i.e., genetically, neurologically, behavioral, etc.). Most importantly, it calls for a stop in linking specific biological or cognitive factors to broad diagnostic (based on the DSM) disorders (Insel and Cuthbert, 2015).

Recently, a converging framework has been put forward that applies such approach to antisocial individuals, including individuals with high levels of psychopathic traits (Brazil et al., 2018). In this bio-cognitive approach, it is suggested to use information from different levels, to form latent categories on which individuals are grouped, that may be better reflect underlying (neurobiological) dysfunctions. Subsequently, these newly defined latent categories may be more effective in guiding interventions and treatment. The approach will use different types of data (i.e., genetics, neuroimaging, cognition) to develop "fingerprints" of individuals that describe that individual based on their unique combination on different dimensions (see **Figure 2**).

Neurophysiology to guide personalized medicine has already been proved to be very promising in another

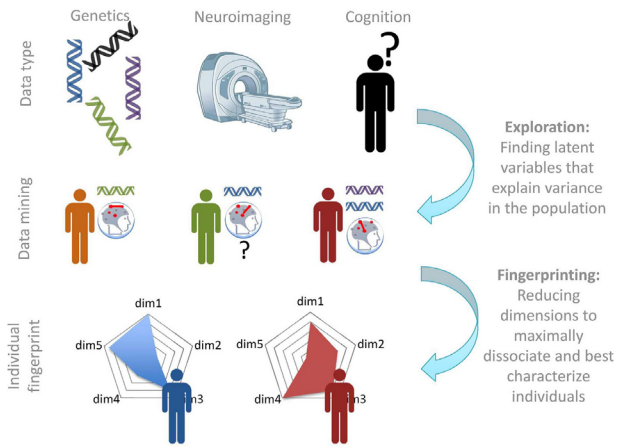


FIGURE 2 | Exploration and fingerprinting of multimodal data. Once the required data are available, various data mining approaches will be required to determine new categories and the measures that describe them. In the second step, the most diagnostic measures can be summarized in a "fingerprint" or profile that can be used to describe each individual on a number of continuous latent dimensions (Brazil et al., 2018).

domain of psychiatry, that of depression. Using data from a consortium, Drysdale et al. (2017) used fMRI connectivity analyses to form "biotypes" on the basis of dysfunctional connectivity patterns. These subtypes of depression were also related to effectiveness of transcranial magnetic stimulation. The authors also pointed to the importance of creating profiles of neurophysiological dysfunction that cross diagnostic boundaries and that can ultimately guide targeted intervention.

However, despite new insights in the complex nature of brain networks (as described in the previous section), there is a lack of studies investigating neural communication within specific frequency bands in psychiatry in general and psychopathic personality more specifically. Moreover, there is a lack of studies that look into dysfunctional topological properties of neural communication within these neural networks (Bullmore and Sporns, 2009). Previous studies are unable to directly evaluate how psychopathy-related connectivity abnormalities actually impact the efficiency and effectiveness of neural information transfer and integration. Also, given the complex structure of psychopathic personality, it is likely that particular traits within psychopathic personality (i.e., more related to F1 or F2 traits, or boldness, meanness, or disinhibition) are differentially associated with complex brain networks in different frequency bands, and with different topological properties of the functional connectivity. In a recent study, Tillem et al. (2018) applied a novel graph theory analysis, minimum spanning tree (MST) analysis, to resting-state EEG data. They found that the interpersonal-affective traits of psychopathy (F1) were associated with decreased efficiency in neural communication between both local and distal brain regions. Conversely, the impulsive-antisocial traits of psychopathy (F2) were associated with increased efficiency of neural communication between both local and distal brain regions.

In my view, the future of an understanding of empathy in psychopathic personality lies with studying the complex networks in the brain in combination with the use of other levels of information (i.e., genetics and cognition). Based on that, profiles of individuals can be formed that can be used to guide neurophysiological informed personalized treatment interventions that ultimately reduce violent transgressions in individuals with psychopathic traits. For example, using brain modulation techniques such as transcranial direct current stimulation (tDCS), activity in particular neural networks can be modulated, thereby modulating its activation and related cognition or behavior in treated individuals. For instance, a study by Choy et al. (2018) showed that when modulating activity with tDCS in the prefrontal cortex, healthy adult individuals were less intended to use aggression during an aggression task. These results point out that tDCS might be a promising alternative treatment for forensic populations (see for example Sergiou et al., 2020).

CONCLUSION

In sum, in this review, the current knowledge on the social neuroscience of empathy in psychopathic personality is discussed, thereby contributing to a better insight in the empathic brain of psychopaths. It is argued that it is important to incorporate data from neuroscience in social sciences, because behavior, especially within the laboratory during experiments, will not reveal the whole picture behind this complex personality. Social neuroscience may unravel differences in functional brain networks that relate to the "empathic brain" of persons with elevated levels of psychopathic personality. Insight in these different complex relations will ultimately lead to a better understanding of this personality and how to target dysfunctional behavior accompanying this personality (e.g., aggression and violence).

To go forward, there is a need for a new approach in studying complex mechanisms, such as empathy, in psychopathic personality. I think that the new way forward must be based on frameworks (e.g., Insel et al., 2010; Brazil et al., 2018) that underscore the need of integration of multiple levels of data types, including neurobiological based information to classify psychopathic personality. By doing so, precision medicine (or personalized medicine; Wium-Andersen et al., 2017) will become a very promising new treatment strategy that can guide social science, including psychology, in developing new and effective interventions for psychopathy.

AUTHOR CONTRIBUTIONS

JD has developed the idea for the review and has written the whole manuscript.

REFERENCES

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., and Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron* 65, 550–562. doi: 10.1016/j.neuron.2010.02.005

Arzy, S., and Danziger, S. (2014). The science of neuropsychiatry: past, present, and future. *J. Neuropsych. Clin. Neurosci.* 26, 392–395. doi: 10.1176/appi.neuropsych.13120371

Baird, A. D., Scheffer, I. E., and Wilson, S. J. (2011). Mirror neuron system involvement in empathy: a critical look at the evidence. *Soc. Neurosci.* 6, 327–335. doi: 10.1080/17470919.2010.547085

Baskin-Sommers, A. R., Curtin, J. J., and Newman, J. P. (2011). Specifying the attentional selection that moderates the fearlessness of psychopathic offenders. *Psychol. Sci.* 22, 226–234. doi: 10.1177/0956797610396227

Batson, C. D. (2009). "These things called empathy: eight related but distinct phenomena," in *The Social Neuroscience of Empathy*, eds J. Decety and W. Ickes (Cambridge: MIT press), 3–15.

Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., Grodd, W., et al. (2005). Deficient fear conditioning in psychopathy: a functional magnetic resonance imaging study. *Arch. Gen. Psychiatry* 62, 799–805.

Bird, G., and Viding, E. (2014). The self to other model of empathy: providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neurosci. Biobehav. Rev.* 47, 520–532. doi: 10.1016/j.neubiorev.2014.09.021

Blair, J. (1996). Theory of mind in the psychopath. *J. For. Psychiatry* 7, 15–25. doi: 10.1080/09585189608409914

Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition* 57, 1–29. doi: 10.1016/0010-0277(95)00676-p

Blair, R. J. R. (2001). Neurocognitive models of aggression, the antisocial personality disorders, and psychopathy. *J. Neurol. Neurosurg. Psychiatry* 71, 727–731. doi: 10.1136/jnnp.71.6.727

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends Cogn. Sci.* 11, 387–392. doi: 10.1016/j.tics.2007.07.003

Blair, R. J. R. (2008). The amygdala and ventromedial prefrontal cortex: functional contributions and dysfunction in psychopathy. *Philos. Trans. R. Soc. Lon. B Biol. Sci.* 363, 2557–2565. doi: 10.1098/rstb.2008.0027

Blair, R. J. R. (2011). Moral judgment and psychopathy. *Emot. Rev.* 3, 296–298. doi: 10.1177/1754073911406297

Blair, R. J. R. (2013). Psychopathy: cognitive and neural dysfunction. *Dialogues Clin. Neurosci.* 15, 181–190.

Blair, R. J. R. (2015a). Psychopathic traits from an RDoC perspective. *Curr. Opin. Neurobiol.* 30, 79–84. doi: 10.1016/j.conb.2014.09.011

Blair, R. J. R. (2015b). Reward processing, functional connectivity, psychopathy and RDoC. *Biol. Psychiatry* 78, 592–593. doi: 10.1016/j.biopsych.2015.08.014

Botvinick, M., Jha, A., Bylsma, L. M., Fabian, S. A., Solomon, P. E., and Prkachin, K. M. (2005). Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain. *NeuroImage* 25, 312–319. doi: 10.1016/j.neuroimage.2004.11.043

Brandle, M., Zhou, H., Smith, B. R. K., Marriott, D., Burke, R., Tabaei, B. P., et al. (2003). The direct medical cost of type 2 diabetes. *Diabetes Care* 26, 2300–2304. doi: 10.2337/diacare.26.8.2300

Brazil, I. A., van Dongen, J. D. M., Maes, J. H. R., Mars, R. B., and Baskin-Sommers, A. R. (2018). Classification and treatment of antisocial individuals: From behavior to biocognition. *Neurosci. Biobehav. Rev.* 91, 259–277. doi: 10.1016/j.neubiorev.2016.10.010

Brook, M., Brieman, C. L., and Kosson, D. S. (2013). Emotion processing in Psychopathy Checklist—assessed psychopathy: a review of the literature. *Clin. Psychol. Rev.* 33, 979–995. doi: 10.1016/j.cpr.2013.07.008

Brook, M., and Kosson, D. S. (2013). Impaired cognitive empathy in criminal psychopathy: evidence from a laboratory measure of empathic accuracy. *J. Abnorma. Psychol.* 122, 156–166. doi: 10.1037/a0030261

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annu. N.Y. Acad. Sci.* 1124, 1–38. doi: 10.1196/annals.1440.011

Bullmore, E., and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 186–198. doi: 10.1038/nrn2575

Bunge, S. A., Dudukovic, N. M., Thomasson, M. E., Vaidya, C. J., and Gabrieli, J. D. E. (2002). Immature frontal lobe contributions to cognitive control in children: evidence from fMRI. *Neuron* 33, 301–311. doi: 10.1016/s0896-6273(01)00583-9

Choy, O., Raine, A., and Hamilton, R. H. (2018). Stimulation of the prefrontal cortex reduces intentions to commit aggression: a randomized, double-blind, placebo-controlled, stratified, parallel-group trial. *J. Neurosci.* 38, 6505–6512. doi: 10.1523/jneurosci.3317-17.2018

Cleckley, H. (1976). *The Mask of Sanity*. St. Louis, Mo: Mosby.

Dawel, A., O'Kearney, R., McKone, E., and Palermo, R. (2012). Not just fear and sadness: meta-analytic evidence of pervasive emotion recognition deficits for facial and vocal expressions in psychopathy. *Neurosci. Biobehav. Rev.* 36, 2288–2304. doi: 10.1016/j.neubiorev.2012.08.006

de Waal, F. B. (2012). The antiquity of empathy. *Science* 336, 874–876. doi: 10.1126/science.1220999

Decety, J. (2010). The neurodevelopment of empathy in humans. *Dev. Neurosci.* 32, 257–267. doi: 10.1159/000317771

Decety, J. (2012). *Empathy: From Bench to Bedside*. Cambridge, MA: MIT Press.

Decety, J., Ben-Ami Bartal, I., Uzefovsky, F., and Knafo-Noam, A. (2016). Empathy as a driver of prosocial behavior: highly conserved neurobehavioral mechanisms across species. *Philos. Trans. R. Soc. Lon. B Biol. Sci.* 371:20150077. doi: 10.1098/rstb.2015.0077

Decety, J., Chen, C., Harenski, C., and Kiehl, K. A. (2013a). An fMRI study of affective perspective taking in individuals with psychopathy: imagining another in pain does not evoke empathy. *Front. Hum. Neurosci.* 7:489. doi: 10.3389/fnhum.2013.00489

Decety, J., Skelly, L. R., and Kiehl, K. A. (2013b). Brain response to empathy-eliciting scenarios involving pain in incarcerated individuals with psychopathy. *JAMA Psychiatry* 70, 638–645.

Decety, J., and Jackson, P. L. (2004). The functional architecture of human empathy. *Behav. Cogn. Neurosci. Rev.* 3, 71–100. doi: 10.1177/1534582304267187

Diamond, A. (2002). "Normal development of prefrontal cortex from birth to young adulthood: cognitive functions, anatomy, and biochemistry," in *Principles of Frontal Lobe Function*, eds D. T. Stuss and R. T. Knight (New York: Oxford University Press).

Dolan, M. C., and Fullam, R. S. (2009). Psychopathy and functional magnetic resonance imaging blood oxygenation level-dependent responses to emotional faces in violent patients with schizophrenia. *Biol. Psychiatry* 66, 570–577. doi: 10.1016/j.biopsych.2009.03.019

Domes, G., Hollerbach, P., Vohs, K., Mokros, A., and Habermeyer, E. (2013). Emotional empathy and psychopathy in offenders: an experimental study. *J. Personal. Disord.* 27, 67–84. doi: 10.1521/pedi.2013.27.1.67

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., et al. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.* 23, 28–38. doi: 10.1038/nm.4246

Eisenberg, N., and Eggum, N. D. (2009). "Empathic responding: sympathy and personal distress," in *The social Neuroscience of Empathy*, eds J. Decety and W. Ickes (Cambridge: MIT press).

Fowles, D. C. (1988). Psychophysiology and psychopathology: a motivational approach. *Psychophysiology* 25, 373–391. doi: 10.1111/j.1469-8986.1988.tb01873.x

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119, 593–609. doi: 10.1093/brain/119.2.593

Gibbon, S., Duggan, C., Stoffers, J., Huband, N., Völlm, B. A., Ferriter, M., et al. (2010). Psychological interventions for antisocial personality disorder. *Cochrane Database Syst. Rev.* 2010:CD007668.

Glenn, A. L., Raine, A., Schug, R. A., Young, L., and Hauser, M. (2009). Increased DLPFC activity during moral decision-making in psychopathy. *Mol. Psychiatry* 14, 909–911. doi: 10.1038/mp.2009.76

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108. doi: 10.1126/science.1062872

Grimm, S., Boesiger, P., Beck, J., Schuepbach, D., Bermpohl, F., Walter, M., et al. (2009). Altered negative BOLD responses in the default-mode network during emotion processing in depressed subjects. *Neuropsychopharmacology* 34, 932–843.

Hare, R. D. (1980). A research scale for the assessment of psychopathy in criminal populations. *Pers. Individ. Diff.* 1, 111–117.

Hare, R. D. (1991). *The Hare Psychopathy Checklist-Revised: Manual*. Toronto: ON: Multi-Health Systems.

Hare, R. D. (1999). *Without Conscience: The Disturbing World of the Psychopaths Among us*. New York, NY: Guilford.

Hare, R. D. (2003). *The Hare Psychopathy Checklist–Revised*. Toronto: ON: Multi-Health Systems.

Harenski, C. L., Harenski, K. A., Shane, M. S., and Kiehl, K. A. (2010). Aberrant neural processing of moral violations in criminal psychopaths. *J. Abnorm. Psychol.* 119, 863–874. doi: 10.1037/a0020979

Harris, G. T., and Rice, M. E. (2006). "Treatment of psychopathy," in *Handbook of Psychopathy*, ed. C. J. Patrick (New York, NY: Guilford), 555–572.

Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and human. *J. Cogn. Neurosci.* 21, 1229–1243. doi: 10.1162/jocn.2009.21189

Hoffman, M. L. (1990). Empathy and justice motivation. *Motiv. Emot.* 14, 151–172. doi: 10.1007/bf00991641

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., and Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biol.* 3:e79. doi: 10.1371/journal.pbio.0030079

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., et al. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am. J. Psychiatry* 167, 748–751. doi: 10.1176/appi.ajp.2010.09091379

Insel, T. R., and Cuthbert, B. N. (2015). Brain disorders? Precisely. *Science* 348, 499–500. doi: 10.1126/science.aab2358

Jackson, P. L., Brunet, E., Meltzoff, A. N., and Decety, J. (2006). Empathy examined through the neural mechanisms involved in imagining how I feel versus how you feel pain. *Neuropsychologia* 44, 752–761. doi: 10.1016/j.neuropsychologia.2005.07.015

Jackson, P. L., Meltzoff, A. N., and Decety, J. (2005). How do we perceive the pain of others: a window into the neural processes involved in empathy. *NeuroImage* 24, 771–779. doi: 10.1016/j.neuroimage.2004.09.006

Jacob, P. (2008). What do mirror neurons contribute to human social cognition? *Mind Lang.* 23, 190–223. doi: 10.1111/j.1468-0017.2007.00337.x

Kajonius, P. J., and Björkman, T. (2020). Individuals with dark traits have the ability but not the disposition to empathize. *Personal. Individ. Differ.* 155:109716. doi: 10.1016/j.paid.2019.109716

Keysers, C., and Gazzola, V. (2006). Towards a unifying neural theory of social cognition. *Progr. Brain Res.* 156, 379–401. doi: 10.1016/s0079-6123(06)56021-2

Keysers, C., and Gazzola, V. (2014). Dissociating the ability and propensity for empathy. *Trends Cogn. Sci.* 18, 163–166. doi: 10.1016/j.tics.2013.12.011

Keysers, C., Kaas, J. H., and Gazzola, V. (2010). Somatosensation in social perception. *Nat. Rev. Neurosci.* 11, 417–428. doi: 10.1038/nrn2833

Kiehl, K. A., and Buckholtz, J. W. (2010). Inside the mind of a psychopathy. *Sci. Am. Mind* 21, 22–29.

Kiehl, K. A., Smith, A. M., Hare, R. D., Mendrek, A., Forster, B. B., Brink, J., et al. (2001). Limbic abnormalities in affective processing by criminal psychopaths as revealed by functional magnetic resonance imaging. *Biol. Psychiatry* 50, 677–684. doi: 10.1016/s0006-3223(01)01222-7

Koenigs, M., Baskin-Sommers, A., Zeier, J., and Newman, J. P. (2011). Investigating the neural correlates of psychopathy: a critical review. *Mol. Psychiatry* 16, 792–799. doi: 10.1038/mp.2010.124

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446, 908–911. doi: 10.1038/nature05631

Kovács, ÁM., Téglás, E., and Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Sci* 330, 1830–1834. doi: 10.1126/science.1190792

Lamm, C., Batson, C. D., and Decety, J. (2007). The neural basis of human empathy: effects of perspective-taking and cognitive appraisal. *Cogn. Neurosci.* 19, 42–58. doi: 10.1162/jocn.2007.19.1.42

Lamm, C., Decety, J., and Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage* 54, 2492–2502. doi: 10.1016/j.neuroimage.2010.10.014

Lamm, C., and Majdandžiæ, J. (2015). The role of shared neural activations, mirror neurons, and morality in empathy–A critical comment. *Neurosci. Res.* 90, 15–24. doi: 10.1016/j.neures.2014.10.008

Levenson, M., Kiehl, K., and Fitzpatrick, C. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *J. Personal. Soc. Psychol.* 68, 151–158. doi: 10.1037/0022-3514.68.1.151

Li, W., Mai, X., and Liu, C. (2014). The default mode network and social understanding of others: what do brain connectivity studies tell us? *Front. Hum. Neurosci.* 8:74. doi: 10.3389/fnhum.2014.00074

Lilienfeld, S. O., and Andrews, B. P. (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal populations. *J. Personal. Assess.* 66, 488–524. doi: 10.1207/s15327752jpa6603_3

Lilienfeld, S. O., and Widows, M. R. (2005). *Psychopathic Personality Inventory–Revised: Professional Manual*. Odessa: PAR.

Lockwood, P. L. (2016). The anatomy of empathy: vicarious experience and disorders of social cognition. *Behav. Brain Res.* 311, 255–266. doi: 10.1016/j.bbr.2016.05.048

Luo, Q., Nakic, M., Wheatley, T., Richell, R., Martin, A., and Blair, R. J. R. (2006). The neural basis of implicit moral attitude—an IAT study using event-related fMRI. *Neuroimage* 30, 1449–1457. doi: 10.1016/j.neuroimage.2005.11.005

Lykken, D. T. (1995). *The Antisocial Personalities*. London: Psychology Press.

Mars, R. B., Neubert, F.-X., Noonan, M. P., Sallet, J., Toni, I., and Rushworth, M. F. S. (2012). On the relationship between the "default mode network" and the "social brain". *Front. Hum. Neurosci.* 6:189. doi: 10.3389/fnhum.2012.00189

Marsh, A. A., and Blair, R. J. R. (2008). Deficits in facial affect recognition among antisocial populations: a meta-analysis. *Neurosci. Biobehav. Rev.* 32, 454–465. doi: 10.1016/j.neubiorev.2007.08.003

Meffert, H., Gazzola, V., den Boer, J. A., Bartels, A. A., and Keyser, C. (2013). Reduced spontaneous but relatively normal deliberate vicarious representations in psychopathy. *Brain* 136, 2550–2562. doi: 10.1093/brain/awt190

Morelli, S. A., Lieberman, M. D., and Zaki, J. (2015). The emerging study of positive empathy. *Soc. Personal. Psychol. Compass* 9, 57–68. doi: 10.1111/spc3.12157

Morrison, I., and Downing, P. E. (2007). Organization of felt and seen pain responses in anterior cingulate cortex. *Neuroimage* 37, 642–651. doi: 10.1016/j.neuroimage.2007.03.079

Morrison, I., Lloyd, D., di Pellegrino, G., and Roberts, N. (2004). Vicarious responses to pain in anterior cingulate cortex: Is empathy a multisensory issue? *Soc. Cogn. Affect. Neurosci.* 4, 270–278. doi: 10.3758/cabn.4.2.270

Morton, J., and Frith, U. (1995). "Causal modeling: a structural approach to developmental psychopathology," in *Series on Personality Processes. Developmental Psychopathology*, Vol. 1, eds D. Cicchetti and D. J. Cohen (Hoboken, NL: John Wiley & Sons), 357–390.

Murphy, F. C., Nimmo-Smith, I. A. N., and Lawrence, A. D. (2003). Functional neuroanatomy of emotions: a meta-analysis. *Cogn. Affect. Behav. Neurosci.* 3, 207–233. doi: 10.3758/cabn.3.3.207

Neumann, C., Hare, R. D., and Newman, J. P. (2007). The super-ordinate nature of the psychopathy checklist-revised. *J. Personal. Disord.* 21, 102–117. doi: 10.1521/pedi.2007.21.2.102

Newman, J. P. (1998). "). Psychopathic behavior: an information processing perspective," in *Psychopathy: Theory, Research and Implications for Society*, eds D. J. Cooke, A. E. Forth, and R. D. Hare (Berlin: Springer Netherlands), 81–104. doi: 10.1007/978-94-011-3965-6_5

Newman, J. P., Curtin, J. J., Bertsch, J. D., and Baskin-Sommers, A. R. (2010). Attention moderates the fearlessness of psychopathic offenders. *Biol. Psychiatry* 67, 66–70. doi: 10.1016/j.biopsych.2009.07.035

Newman, J. P., Patterson, C. M., and Kosson, D. S. (1987). Response perseveration in psychopaths. *J. Abnorm. Psychol.* 96, 145–148. doi: 10.1037/0021-843x.96.2.145

Onishi, K. H., and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science* 308, 255–258. doi: 10.1126/science.1107621

Patrick, C. J., Cuthbert, B. N., and Lang, P. J. (1994). Emotion in the criminal psychopath: fear image processing. *J. Abnorm. Psychol.* 103, 523–534. doi: 10.1037/0021-843x.103.3.523

Patrick, C. J., Fowles, D. C., and Krueger, R. F. (2009). Triarchic conceptualization of psychopathy: developmental origins of disinhibition, boldness, and meanness. *Dev. Psychopathol.* 21, 913–938. doi: 10.1017/s0954579409000492

Paulhus, D. L., Neumann, C. S., and Hare, R. D. (2016). *Manual for the Self-reported Psychopathy Scale*, 4th Edn. Toronto: Multi-Health Systems.

Preston, S. D., Bechara, A., Damasio, H., Grabowski, T. J., Stansfield, R. B., Mehta, S., et al. (2007). The neural substrates of cognitive empathy. *Soc. Neurosci.* 2, 254–275. doi: 10.1080/17470910701376902

Preston, S. D., and de Waal, F. B. M. (2002). Empathy: its ultimate and proximate bases. *Behav. Brain Scence* 25, 1–72.

Price, D. D. (2000). Psychological and neural mechanisms of the affective dimension of pain. *Science* 288, 1769–1772. doi: 10.1126/science.288.5472.1769

Raichle, M. E. (2015). The restless brain: how intrinsic activity organizes brain function. *Philos. Trans. R. Soc. Lon. B Biol. Sci.* 370:20140172. doi: 10.1098/rstb.2014.0172

Richell, R. A., Mitchell, D. G., Newman, C., Leonard, A., Baron-Cohen, S., and Blair, R. J. (2003). Theory of mind and psychopathy: can psychopathic individuals read the 'language of the eyes'? *Neuropsychologia* 41, 523–526. doi: 10.1016/s0028-3932(02)00175-6

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Ann. Rev. Neurosci.* 27, 169–192.

Robinson, E. V., and Rogers, R. (2015). Empathy faking in psychopathic offenders: the vulnerability of empathy measures. *J,. Psychopathol. Behav. Assessment* 37, 545–552. doi: 10.1007/s10862-015-9479-9

Rothemund, Y., Ziegler, S., Hermann, C., Gruesser, S. M., Foell, J., Patrick, C. J., et al. (2012). Fear conditioning in psychopaths: event-related potentials and peripheral measures. *Biol. Psychol.* 90, 50–59. doi: 10.1016/j.biopsycho.2012.02.011

Roy, S., Vize, C., Uzieblo, K., van Dongen, J. D. M., Miller, J., Lynam, D., et al. (2020). Triarchic or Septarchic?–Uncovering the Triarchic Psychopathy Measure's (TriPM) Structure. *Personal Disord. Theory, Res. and Treat*. doi: 10.1037/per0000392 [Epub ahead of print].

Ruby, P., and Decety, J. (2004). How would you feel versus how do you think she would feel? A neuroimaging study of perspective taking with social emotions. *J. Cogn. Neurosci.* 16, 988–999. doi: 10.1162/0898929041502661

Salekin, R. T., Worley, C., and Grimes, R. D. (2010). Treatment of psychopathy: a review and brief introduction to the mental model approach for psychopathy. *Behav. Sci. Law* 8, 235–266. doi: 10.1002/bsl.928

Sandoval, A. M. R., Hancock, D., Poythress, N., Edens, J. F., and Lilienfeld, S. (2000). Construct validity of the psychopathic personality inventory in a correctional sample. *J. Personal. Assess.* 74, 262–281. doi: 10.1207/s15327752jpa7402_7

Sellbom, M., and Phillips, T. R. (2013). An examination of the triarchic conceptualization of psychopathy in incarcerated and nonincarcerated samples. *J. Abnorm. Psychol.* 122, 208–214. doi: 10.1037/a0029306

Sergiou, C. S., Woods, A. J., Franken, I. H. A., and van Dongen, J. D. M. (2020). Transcranial direct current stimulation (tDCS) as an intervention to improve empathic abilities and reduce violent behavior in forensic offenders: study protocol for a randomized controlled trial. *Trials* 21, 1–14. doi: 10.1186/s13063-020-4074-0

Shamay-Tsoory, S. G., Harari, H., Aharon-Peretz, J., and Levkovitz, Y. (2010). The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex* 46, 668–677.

Singer, T., and Lamm, C. (2009). The social neuroscience of empathy. *Ann. N.Y. Acad. Sci.* 1156, 81–96.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., and Frith, C. D. (2004). Empathy for pain involves the affective but not the sensory components of pain. *Science* 303, 1157–1161.

Skeem, J. L., Poythress, N., Edens, J. F., Lilienfeld, S. O., and Cale, E. M. (2003). Psychopathic personality or personalities? Exploring potential variants of psychopathy and their implications for risk assessment. *Aggress. Violent Behav.* 8, 513–546.

Sommerville, J. A., and Decety, J. (2006). Weaving the fabric of social interaction: Articulating developmental psychology and cognitive neuroscience in the domain of motor cognition. *Psychon. Bull. Rev.* 13, 179–200.

Stanley, J. H., Wygant, D. B., and Sellbom, M. (2013). Elaborating on the construct validity of the triarchic psychopathy measure in a criminal offender sample. *J. Personal. Assess.* 95, 343–350.

Tassy, S., Deruelle, C., Mancini, J., Leistedt, S., and Wicker, B. (2013). High levels of psychopathic traits alters moral choice but not moral judgment. *Front. Hum. Neurosci.* 7:229. doi: 10.3389/fnhum.2013.00229

Tillem, S., van Dongen, J. D. M., Brazil, I. A., and Baskin-Sommers, A. (2018). Psychopathic traits are differentially associated with efficiency of neural communication. *Psychophysiology* 55:e13194.

van Dongen, J. D. M., Brazil, I. A., van der Veen, F. M., and Franken, I. H. A. (2018). Electrophysiological correlates of empathic processing and its relation to psychopathic meanness. *Cogn. Affect. Behav. Neurosci.* 32, 996–1006.

Wium-Andersen, I. K., Vinberg, M., Kessing, L. V., and McIntyre, R. S. (2017). Personalized medicine in psychiatry. *Nordic J. Psychiatry* 71, 12–19.

Zaki, J., and Ochsner, K. N. (2012). The neuroscience of empathy: progress, pitfalls and promise. *Nat. Neurosci.* 15, 675–680.

Zaki, J., Ochsner, K. N., Hanelin, J., Wager, T. D., and Mackey, S. C. (2007). Different circuits for different pain: patterns of functional connectivity reveal distinct networks for processing pain in self and others. *Soc. Neurosci.* 2, 276–291.

Zaki, J., Wager, T. D., Singer, T., Keysers, C., and Gazzola, V. (2016). The anatomy of suffering: understanding the relationship between nociceptive and empathic pain. *Trends Cogn. Sci.* 20, 249–259.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

GENERAL COMMENTARY
published: 08 January 2020
doi: 10.3389/fpsyg.2019.02880

Commentary: The moral bioenhancement of psychopaths

Elisabetta Sirgiovanni[1]* *and Mirko Daniel Garasic*[2]

[1] Department of Molecular Medicine, Faculty of Pharmacy and Medicine, Sapienza University of Rome, Rome, Italy, [2] Libera Università Maria SS. Assunta, Rome, Italy

Keywords: moral bioenhancement, psychopathy, consent and refusal to treatment, involuntary treatment, open justification

A Commentary on

The moral bioenhancement of psychopaths
by Baccarini E., and Malatesti L. (2017). J. Med. Ethics 43, 697–701. doi: 10.1136/medethics-2016-103537

Baccarini and Malatesti (2017) defend the idea that we must use coercively biomedical means to enhance the morality of a specific group of individuals: psychopaths, diagnosed through the Psychopathy Checklist-Revised (PCL-R) standards (Hare, 2003). Their argument is theoretical, thus it goes independently from the actual effectiveness of existent treatments, and it is based on a logical reasoning. Moral bioenhancement (MB) means include psychotropic drugs, brain stimulations, neurosurgeries, genetic editing, etc.

In short, the authors apply Gerald Gaus' account of *open justification* (Gaus, 1996, 2011), according to which "a prescription addressed to an agent is a reasoning that includes premises that consider the system of reasons (such as beliefs, preferences, etc.) of that agent" (Baccarini and Malatesti, 2017, p. 1). In their view, coercive MB of psychopaths is morally sound and deducible by reasons within the psychopath's cognitive-affective system—even if the psychopath needs not to be able to consciously or sincerely endorse them.

Notoriously psychopaths have Machiavellian traits, a dimension in the Dark Triad (Paulhus and Williams, 2002), including anti-sociality and narcissism. In order to exploit others, the psychopath wishes to live in a society where everyone is cooperative except herself. Consequentially, the psychopath would prescribe MB to other psychopaths. The authors state that an agent must apply to herself a prescription she would accept for others, "if she shares with them the relevant characteristics" (i.e., psychopathic traits), and "unless (s)he can justify to others that the two cases are relevantly different" (Baccarini and Malatesti, 2017, p. 3). Since the psychopath possesses the same personality traits of other psychopaths, the authors claim we would be justified, in Kantian terms, to universalize the prescription of mandatory MB to her.

We believe that this argument is flawed. In sum, we argue that the psychopath's cognitive-affective system would consistently justify reasons against mandatory MB to herself, even if she wishes differently for others, and that the prescription cannot be extended. What "immoral rule" is the best deducible from the psychopath's cognitive-affective system? If we think of human morality as cooperation in evolutionary terms (Curry, 2016), as the authors do, it seems that psychopaths contradict what has been held inter-culturally as a guiding principle of reciprocity, the Golden Rule. On the contrary, psychopaths respond to what we may call, from the triad, a Dark Rule. Psychopaths believe and feel that "one can treat others (i.e., manipulating, hurting, torturing, killing, etc.) in ways that one would not like to be treated." In fact, there is no evidence that psychopaths wish to be treated (even unconsciously) in the same ways they treat others. Research shows that when viewing stimuli depicting bodily injuries adopting an image-self perspective, psychopaths

OPEN ACCESS

Edited by:
José M. Muñoz, Universidad Europea de Valencia, Spain

Reviewed by:
Ezequiel Norberto Mercurio, University of Buenos Aires, Argentina

***Correspondence:**
Elisabetta Sirgiovanni
elisabetta.sirgiovanni@uniroma1.it

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *19 August 2019*
Accepted: *04 December 2019*
Published: *08 January 2020*

Citation:
Sirgiovanni E and Garasic MD (2020) Commentary: The moral bioenhancement of psychopaths. Front. Psychol. 10:2880. doi: 10.3389/fpsyg.2019.02880

have normal neural responses for pain (Decety et al., 2013). These responses do not match the atypical patterns of brain activation psychopaths show when adopting an other-perspective. Thus, the psychopath can consistently justify within her cognitive-affective system that her own case and the other psychopaths' case are relevantly different.

It could be objected that a Dark Rule entails for the psychopaths to accept to be treated by others in ways they do not like to be treated. Yet, we should keep in mind that, for a Kantian, the Dark Rule (i.e., treating others as a means) is intrinsically unethical, hence it is *not* a universalizable rule.

Having pointed out this unconvincing dimension of Baccarini and Malatesti's account, we wish to next raise objections about forcing MB on psychopaths even if that was indeed the case.

Involuntary treatment has been justified by combining public reasons of social security (Persson and Savulescu, 2012, 2019) with other criteria implemented in different legislations (Saya et al., 2019), such as mental incapacity and non-intrusiveness of the treatment. Remarkably, all these criteria are now challenged by recent international standards for the rights of persons with disabilities, where informed consent to mental health services has been vigorously supported in any case (see United Nations, 2006, art. 14; United Nations, 2008, par. 64–65; United Nations, 2019).

With regard to MB of psychopaths, it is questionable that these criteria can be met.

In most cases, it is doubtful to claim that the psychopath's volition is harmed. Remarkably, psychopaths are multifaceted in decision-making, by mainly lacking emotional engagement in moral choice/action while their rational judgment is unimpaired (Cima et al., 2010; Aharoni et al., 2014; Jurjako and Malatesti, 2016). Evolutionists do not see psychopathic traits as expression of an underlying dysfunction, but as a persisting adaptation to certain environments (Glenn et al., 2011). Notably, there are still discrepancies between the PCL-R construct of psychopathy and the corresponding official category of antisocial personality disorder (ASPD) in the DSM (Few et al., 2015). These considerations together could reinforce the argument that we are not totally entitled to classify psychopathy as a proper mental incapacitation. It must be noticed that PCL-R diagnoses are over-inclusive, since the scale attributes psychopathic traits dimensionally to a large group of people, including non-offending and subclinical individuals such as businessmen, lawyers, actors, politicians, and rebels of various sort, not only serial killers and recidivist offenders (Skeem et al., 2011).

Most importantly, MB is far from being the least restrictive or intrusive treatment. This might exclude most MB means, especially those that are irreversible (e.g., neurosurgeries), impact severely on intertwined functions (e.g., psychotropic drugs, brain stimulations, etc.), and that pass on through generations unpredictably (i.e., gene editing).

Moreover, the call for involuntary treatment is not as neutral and objective as often depicted by its promoters (Garasic, 2013). The "greater good for society" behind the suspension of human rights is often charged with biopolitical values, and it exploits the patient/prisoner as a tool to reinforce or instill specific norms/standards in the society. The defense of coercive MB hides an idea of "moral perfectionism" (Cavell, 1990), according to which we must conform to an idealistic and demanding account of morality where moral imperfections or differences are never tolerated and need to be eliminated. Defining the "morally perfect" is a challenge as much as concluding that a society without moral defects would be a better society. What is the prototypical "moral individual" into whom we should transform the psychopath?

This approach creates substantial frictions with the individual rights. For its *moral* specificity, coercive MB interferes tremendously with individual autonomy and freedom without empowering moral competence (Harris, 2011, 2016; Corbellini and Sirgiovanni, 2015). Personal preferences/options belong to a larger spectrum of moral acceptability than that conventionally prescribed by society in a given historical time.

Furthermore, it is unclear whether we should prescribe mandatory MB also to non-psychopathic offenders and preventively to non-offending or subclinical psychopaths. The same reasons of social security, in fact, seem to predispose ourselves (and society) to large extensions of the legitimacy of MB.

In conclusion, we defend the view that the right to refuse MB must be protected. It seems that without consent, psychopathic offenders' incarceration or admission to psychiatric facility are still the only acceptable security measures.

AUTHOR CONTRIBUTIONS

The authors discussed, reviewed, and approved together the entire manuscript. ES conceived and wrote the first part of the manuscript and half of the second part, MG conceived and wrote the second part.

REFERENCES

Aharoni, E., Sinnott-Armstrong, W., and Kiehl, K. A. (2014). What's wrong? Moral understanding in psychopathic offenders. *J. Res. Pers.* 53, 175–181. doi: 10.1016/j.jrp.2014.10.002

Baccarini, E., and Malatesti, L. (2017). The moral bioenhancement of psychopaths. *J. Med. Ethics* 43, 697–701. doi: 10.1136/medethics-2016-103537

Cavell, S. (1990). *Conditions Handsome and Unhandsome: The Constitution of Emersonian Perfectionism*. Chicago, IL: University of Chicago Press.

Cima, M., Tonnaer, F., and Hauser, M. D. (2010). Psychopaths know right from wrong but don't care. *Soc. Cogn. Affect. Neurosci.* 5, 59–67. doi: 10.1093/scan/nsp051

Corbellini, G., and Sirgiovanni, E. (2015). Against paternalistic views on neuroenhancement: a libertarian evolutionary account. *Med. Secoli* 27, 1089–1110. doi: 10.1007/s11299-016-0188-1

Curry, O. S. (2016). "Morality as cooperation: a problem-centred approach," in *The Evolution of Morality,* eds K. Shackelford, and R. D. Hansen (Basel: Springer International Publishing), 27–51. doi: 10.1007/978-3-319-19671-8_2

Decety, J., Chen, C., Harenski, C., and Kiehl, K. A. (2013). An fMRI study of affective perspective taking in individuals with psychopathy: imagining another in pain does not evoke empathy. *Front. Hum. Neurosci.* 7:489. doi: 10.3389/fnhum.2013.00489

Few, L. R., Lynam, D. R., Maples, J. L., MacKillop, J., and Miller, J. D. (2015). Comparing the utility of DSM-5 section II and III antisocial

personality disorder diagnostic approaches for capturing psychopathic traits. *Pers. Disord. Theor. Res. Treat.* 6, 64–74. doi: 10.1037/per0000096

Garasic, M. D. (2013). The singleton case: enforcing medical treatment to put a person to death. *Med. Health Care Philos.* 16, 795–806. doi: 10.1007/s11019-013-9462-8

Gaus, G. (1996). *Justificatory Liberalism: An Essay on Epistemology and Political Theory*. Oxford: Oxford University Press.

Gaus, G. (2011). *The Order of Public Reason: a Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge: Cambridge University Press.

Glenn, A. L., Kurzban, R., and Raine, A. (2011). Evolutionary Theory and Psychopathy. *Aggress. Violent Behav.* 16, 371–380. doi: 10.1016/j.avb.2011.03.009.

Hare, R. (2003). *Hare Psychopathy Checklist-Revised (PCL-R), 2nd Edn*. Toronto, ON: Multi-Health System.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics* 25, 102–111. doi: 10.1111/j.1467-8519.2010.01854

Harris, J. (2016). Moral blindness – The gift of the god machine. *Neuroethics* 9, 269–273. doi: 10.1007/s12152-016-9272-9

Jurjako, M., and Malatesti, L. (2016). Instrumental rationality in psychopathy: implications from learning tasks. *Philos. Psychol.* 29, 717–731. doi: 10.1080/09515089.2016.1144876

Paulhus, D. L., and Williams, K. M. (2002). The Dark Triad of personality: narcissism, Machiavellianism and psychopathy. *J. Res. Pers.* 36, 556–563. doi: 10.1016/S0092-6566(02)00505-6

Persson, I., and Savulescu, J. (2012). *Unfit for the Future*. Oxford: Oxford University Press.

Persson, I., and Savulescu, J. (2019). The duty to be morally enhanced. *Topoi* 38, 7–14. doi: 10.1007/s11245-017-9475-7

Saya, A., Brugnoli, C., Piazzi, G., Liberato, D., Di Ciaccia, G., Niolu, C., et al. (2019). Criteria, procedures, and future prospects of involuntary treatment in psychiatry around the world: a narrative review. *Front. Psychiatry* 10:271. doi: 10.3389/fpsyt.2019.00271

Skeem, J. L., Polaschek, D. L. L., Patrick, C. J., and Lilienfeld, S. O. (2011). Psychopathic personality: bridging the gap between scientific evidence and public policy. *Psychol. Sci. Public Interest* 12, 95–162. doi: 10.1177/1529100611426706

United Nations (2006). *Convention on the Rights of Persons with Disabilities and Its Optional Protocol*. Adopted by UN General Assembly Resolution A/RES/61/106 of 3 December 2006. New York, NY: United Nations. Available online at: https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html

United Nations (2008). *Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment*. Adopted by UN General Assembly Resolution A/63/175 of 28 July 2008. New York, NY: United Nations. Available online at: https://spinternet.ohchr.org/SP/Resolutions/Shared%20Documents/RES/A_HRC_RES_8_8_E.pdf.

United Nations (2019). *Report of the Special Rapporteur on the Rights of Persons with Disabilities*. Adopted by UN General Assembly Resolution A/74/186 of 17 July 2019. New York, NY: United Nations. Available online at: https://www.un.org/ga/search/view_doc.asp?symbol=A/74/186

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 16 October 2019
doi: 10.3389/fpsyg.2019.02343

Neuroscientific and Genetic Evidence in Criminal Cases: A Double-Edged Sword in Germany but Not in the United States?

*Daniela Guillen Gonzalez[1], Merlin Bittlinger[2], Susanne Erk[1] and Sabine Müller[1]**

[1] Research Division of Mind and Brain Research, Department of Psychiatry and Psychotherapy CCM, Berlin Institute of Health, Humboldt University of Berlin, Corporate Member of the Free University of Berlin, Charité – Berlin University of Medicine, Berlin, Germany, [2] QUEST – Center for Transforming Biomedical Research, Berlin Institute of Health, Berlin, Germany

OPEN ACCESS

Edited by:
Elena Rusconi,
University of Trento, Italy

Reviewed by:
Sara Dellantonio,
University of Trento, Italy
Heidi Lene Maibom,
University of Cincinnati, United States

***Correspondence:**
Sabine Müller
mueller.sabine@charite.de

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: 08 July 2019
Accepted: 01 October 2019
Published: 16 October 2019

Citation:
Guillen Gonzalez D, Bittlinger M, Erk S and Müller S (2019) Neuroscientific and Genetic Evidence in Criminal Cases: A Double-Edged Sword in Germany but Not in the United States? Front. Psychol. 10:2343. doi: 10.3389/fpsyg.2019.02343

Aim of the Study: The study examines how neurobiological and genetic explanations of psychopathy influence decision-making of German law students about legal and moral responsibility and sentencing of a defendant in a case of manslaughter. Previous studies from the United States and Germany have been criticized because they partly contradict legal analyses of real-world criminal cases. With a modified design, which integrates the main criticism, we re-examined the impact of biological explanations for psychopathy on decision-making in the courtroom.

Methods: We developed an improved quasi-experimental design to probe three case vignettes presenting different explanations of psychopathy in a criminal case of manslaughter. All three vignettes present the same information about a forensic expert's testimony that is said to report compelling evidence for the diagnosis of "psychopathy." The independent variable being manipulated is the type of information supporting the expert diagnosis: either no biological explanation of "psychopathy" versus a neurological explanation (brain injury) versus a genetic explanation (MAOA gene). The outcome measure is a questionnaire on legal and moral responsibility, free will, the type of custody, and the duration of the sentence. The study is adequately powered. We openly publish the data and all statistical analyses as reproducible R scripts.

Results: The answers of German law students (n = 317) indicate that the omission of a neurobiological explanation is significantly associated with higher ratings of legal responsibility while compared to no biological explanation. However, there was no significant difference on the prison sentencing and type of custody assigned. Furthermore, there was no difference in the self-reported impact of the explanation of psychopathy on the participants' decision-making.

Conclusion: Our findings from German law students corroborates previous research on German judges but is markedly distinct from studies on United States judges. Whereas in the United States, biological information seems to have a mitigating effect, it seems to increase the rate of involuntary commitment to forensic psychiatric hospitals in Germany.

Keywords: neurolaw, neuroscience evidence, responsibility, culpability, psychopathy

INTRODUCTION

Both United States courts and commentators have discussed the use of neuroscientific and genetic evidence in criminal cases as a "double-edged sword" for the defendant (Denno, 2012). On the one side, such evidence has a mitigating potential because it reduces the culpability of the defendant. On the other side, it can be an aggravating factor because it supports the assumption of future dangerousness. However, the United States legal theorist Denno (2015), who has analyzed hundreds real criminal cases, in which biological evidence was introduced, calls the double-edged sword theory a myth.

Indeed, neuroscience and genetic evidence is increasingly being introduced in criminal cases in the United States (Denno, 2012, 2015; Denno and McGivney, 2013; Farahany, 2015), in Canada (Chandler, 2015), Western Europe (Catley and Claydon, 2015; De Kogel and Westgeest, 2015), and Australia (Alimardani and Chin, 2018). In most of these cases, clinically established techniques such as EEG, structural brain imaging, and positron emission tomography have been used to demonstrate brain damage, whereas fMRI and neurogenetics have been used only in few cases (Fuss, 2016).

This paper contributes to the debate about the double-edged sword theory. First, we review the debate about the nature and the causes of psychopathy, and discuss its particular importance for criminal justice. Then we summarize the results of experimental studies investigating the double-edged sword effect.

The main part of the paper presents the results of our own experimental study that has investigated how neurobiological and genetic explanations of psychopathy influence the decision-making of German law students about legal and moral responsibility and sentencing of a defendant in a case of manslaughter. Our own study is based on older studies, but we have modified the design in order to integrate the main criticism of these studies.

Finally, we discuss the reasons for the inconsistent results of the different studies in the light of studies which have comprehensively analyzed real criminal cases in different countries. We suggest that the question whether neuroscientific and genetic evidence in criminal cases is a double-edged sword cannot be answered in general. Rather, the answer depends strongly on the system of criminal justice of a given country.

Abbreviations: α, level of significance of the probability value; χ^2, chi-squared; η_p^2, partial eta-squared; df, degrees of freedom; M, mean; MAOA, monoamine oxidase A; n, number of participants for a given subset of the sample; N, total number of participants in the sample; OFC/VLPFC, orbitofrontal plus ventrolateral prefrontal cortex; SD, standard deviation; vmPFC, ventromedial prefrontal cortex.

The Nature and the Causes of Psychopathy and Its Particular Importance for Criminal Justice

Psychopathy is in the focus of neuroscientific and genetic research, although after a long and controversial debate (Crego and Widiger, 2015), it was not included as a stand-alone personality disorder in the DSM-5 (American Psychiatric Association, 2013).

The focus on psychopathy is justified because psychopathy is "one of the strongest dispositional predictors of aggression and violence" (Reidy et al., 2015). Psychopaths commit the most severe acts of violence; they commit twice as many violent crimes as non-psychopathic offenders and their risk of violent recidivism is at least five times higher (Reidy et al., 2015).

According to the influential psychopathy researcher Hare (1996, p. 25), psychopathy is "a devastating disorder defined by a constellation of affective, interpersonal and behavioral characteristics, including egocentricity; impulsivity; irresponsibility; shallow emotions; lack of empathy, guilt or remorse; pathological lying; manipulativeness; and the persistent violation of social norms and expectations." Hare (1996, p. 26) describes psychopaths as "intraspecies predators who use charm, manipulation, intimidation, and violence to control others and to satisfy their own selfish needs. Lacking in conscience and in feelings for others, they cold-bloodedly take what they want and do as they please, violating social rules without the slightest sense of guilt or regret."

For the diagnosis of psychopathy, forensic psychiatrists mostly use the Hare (2003) Psychopathy Checklist-Revised (PDL-R) and its derivatives (Reidy et al., 2015).

There are two opposing perspectives on psychopathy: (1) psychopathy is a mental disorder based on structural and functional dysfunctions of several brain areas, and (2) the developmental form of psychopathy is a moral or social disorder, but not a biological disorder.

Blair (2013) promotes the first perspective by describing psychopathy as a developmental disorder characterized by pronounced emotional deficits marked by reduction in guilt and empathy, and increased risk for displaying antisocial behavior. Blair (2013) emphasizes that psychopathy is not equivalent to antisocial personality disorder from the diagnostic systems DSM-IV-R or ICD-10, which focus on the antisocial behavior rather than underlying causes, i.e., the emotion dysfunction. Blair (2013) has suggested that the emotion dysfunction relates to three core functional impairments: the association of stimuli with reinforcement, the representation of expected value information and prediction error signaling. He hypothesizes that these functional impairments relate to the observed dysfunction seen

in structural and functional MRI studies within the amygdala, vmPFC, and (only in youth populations) striatum (Blair, 2013).

Reimer (2008) suggests describing psychopathy without the language of disorder. According to an evolutionary model, psychopathy represents an alternative genetic strategy that is successful only at a particular low relative frequency in the population (Reimer, 2008). This idea is supported by game-theoretical models of non-cooperators who move between groups and "prey" on naïve cooperators (Dugatkin, 1992). This idea has been elaborated in sociobiology. Mealey (1995) explained sociopathy as "the expression of a frequency-dependent life strategy which is selected, in dynamic equilibrium, in response to certain varying environmental circumstances." Reimer (2008) suggests that psychopaths are not disordered in any biological sense, but only different from the majority of people. Psychopaths are not impaired, but especially capable. They have a "pro-individual personality" with special capacities for "successful individualization" (Reimer, 2008). Particularly, they are capable of ignoring the distress of others and are better able to resist attempts at "moral" social reinforcing (Reimer, 2008). With regard to the amygdala-dysfunction theory of psychopathy, Reimer (2008) does not deny the role of the amygdala. Rather she says that the special development of the amygdala enables the "pro-individual personality" to successfully pursue the person's goals, including reproductive ones, "without the hindrances imposed by other regarding norms" (Reimer, 2008). In this way, the "pro-individual personality" is able to insure the dissemination of her pro-individual genes in future generations (Reimer, 2008).

The view that psychopathy is a moral disorder that is not caused by a lack of capacities is supported by a study suggesting that psychopaths do understand the distinction between right and wrong, but do not care about such knowledge or the consequences that ensue from their morally inappropriate behavior (Cima et al., 2010).

Particularly the fact that many psychopaths are successful supports Reimer's suggestion to describe psychopathy without the language of disorder. Babiak and Hare (2006) found a higher rate of psychopaths in the business world than in the general population (3.5% vs. 0.6–1%). Although both successful (not incarcerated) and unsuccessful (incarcerated) psychopaths show autonomic hyporeactivity (low resting heart rate), reduced emotional empathy, risky decision making and sensation-seeking, the successful psychopaths seem to have intact or even enhanced neurobiological functioning, which enables them to lie, con and manipulate successfully (Gao and Raine, 2010). In contrast, unsuccessful psychopaths have more cognitive and emotional deficits and tend to violent offending instead of white collar criminality (Gao and Raine, 2010).

In 1996, Hare (1996) noted that in most jurisdictions, psychopathy is considered an aggravating rather than a mitigating factor in determining criminal responsibility. However, research evidence explaining psychopathy in terms of an affective deficit, a thought disorder or brain dysfunction might lead some to view psychopathy as a mitigating factor (Hare, 1996). Hare (1996) considers a psychiatrist's speculation that psychopathy would perhaps become "the kiss of life rather than the kiss of death" in first-degree murder cases, as "appalling, because psychopaths are calculating predators whose behavior must be judged by the rules of the society in which they live."

The causes of psychopathy are controversial. Early studies investigated correlations between physiological indices such as heart rate and electrodermal activity with aggression, psychopathy/sociopathy, and conduct problems (Lorber, 2004). Low autonomic activity might contribute to the development of antisocial and criminal behavior, because it is a marker for fearlessness, and leads to sensation-seeking behavior (Raine, 2002).

Prenatal factors also contribute to antisocial and violent behavior, particularly pregnancy complications, birth complications, maternal smoking and alcohol consume during pregnancy; these factors strongly interact with each other (Raine, 2002).

Current research concentrates on the neurotransmitters serotonin, dopamine and vasopressin, the steroid hormones testosterone and cortisol, and brain structure and function (Rosell and Siever, 2015). Particularly, the amygdala, the prefrontal cortex and the striatum are in the focus of research (Rosell and Siever, 2015). However, the phenomenological heterogeneity of aggression is a source of inconsistencies between studies, and the categorical nature of psychiatric diagnoses is another critical issue (Rosell and Siever, 2015).

Sociopathy or chronic antisocial behavior can be a developmental or an acquired disorder (Mendez, 2009). The most famous case of acquired sociopathy caused by brain injury certainly is Phineas Gage. This case has become a scientific myth, perhaps because it is fascinating to watch someone break bad (Kean, 2014). Focal lesions affecting vmPFC and adjacent OFC/VLPFC include strokes, trauma, tumors, infections, and a ruptured anterior commissure aneurysm, and can lead to alterations in social and moral behavior (Mendez, 2009).

A meta-analysis of 43 structural and functional imaging studies showed significantly reduced prefrontal structure and function in antisocial individuals (Yang and Raine, 2009). A study with 56 males showed that men with lower amygdala volume exhibited higher levels of aggression and psychopathic features from childhood to adulthood (Pardini et al., 2014).

A systematic mapping of lesions with known temporal association to criminal behavior has revealed that the lesion sites are spatially heterogenous, including the medial prefrontal cortex and the orbitofrontal cortex. However, all these lesions are part of a unique functionally connected brain network, which is involved in moral decision making (Darby et al., 2018).

Evidence from behavioral genetics supports the conclusion that a significant amount of the variance in antisocial personality is due to genetic contributions. A meta-analytic review on behavioral genetic etiological studies of antisocial personality and behavior showed that 56% of the variance of antisocial personality and behavior can be explained through genetic influences, with 11% due to shared non-genetic influences and 31% due to unique non-genetic influences (Ferguson, 2010).

Particularly prominent is the MAOA gene, which is located on the X-chromosome. It encodes the MAOA enzyme, which metabolizes norepinephrine, serotonin and dopamine

(Caspi et al., 2002). In males, a point mutation in the MAOA gene, which causes a complete MAOA deficiency, is associated with abnormal aggressive behavior and impulsivity in a large Dutch kindred (Brunner et al., 1993).

Caspi et al. (2002) found in males a gene × environment interaction between the MAOA gene and childhood maltreatment. Maltreated male children with high MAOA activity were significantly less likely to develop child conduct disorder, a disposition toward violence, an adult antisocial personality disorder and convictions for violent offenses (Caspi et al., 2002). Although the low-MAOA genotype on its own did not significantly increase the risk of developing antisocial behavior, it increased the risk for developing antisocial behavior among males who suffered maltreatment (Caspi et al., 2002).

Another research group replicated the results of Caspi's study through the investigation of another sample of boys and a meta-analysis (Kim-Cohen et al., 2006).

A Finnish prisoner study with over 500 offenders revealed that a MAOA low-activity genotype and the CDH13 gene are associated with severe recidivistic violent behavior (Tiihonen et al., 2015).

However, a recent systematic meta-analysis did not find any significant association between any polymorphism analyzed, and aggression and violence; even subgroup analyses did not show any consistent findings (Vassos et al., 2014). Since no gene of major effect for aggression has been identified, the authors of the meta-analysis consider any approach to use genetic markers for risk prediction or to mitigate criminal responsibility questionable (Vassos et al., 2014). Tiihonen and coauthors emphasize, too, that the sensitivity and specificity of the genotype findings are much too low for any screening purposes for prevention of violent offending, and that putative risk factors such as genotype do not have a legal role in judgment about offenders (Tiihonen et al., 2015).

The relationship between genes and aggressive and antisocial behavior is much more complex than formerly believed. On the one hand, behavioral genetics shows that distinct polymorphisms of genes, which code for proteins controlling neurotransmitter function, are associated with individual vulnerability to aversive experiences, and may result in an increased risk of developing psychopathologies associated with violence (Palumbo et al., 2018). On the other hand, epigenetic studies indicate that aversive experiences particularly during prenatal life, infancy and early adolescence can introduce lasting epigenetic marks in genes, thus favoring the emergence of dysfunctional behaviors, including exaggerated aggression (Palumbo et al., 2018).

In the development of violent behavior and aggression, biological, psychodynamic and social factors play a role (Sopromazde and Tsiskaridze, 2018). Social and biological factors do not have simply an additive effect; rather the presence of both factors exponentially increases the rates of antisocial and violent behavior (Raine, 2002). In a good social environment, the association between biological factors and antisocial behavior is stronger (Raine, 2002).

Maltreatment during childhood and maternal withdrawal in infancy are significantly associated with antisocial personality disorder (Shi et al., 2012). The Cambridge Study in Delinquent Development, a prospective longitudinal study, which started in 1961, suggested that "the best predictors of psychopathy" were "having a convicted parent, physical neglect, low paternal involvement, low family income, and coming from a disrupted family" (Reidy et al., 2015). The transmission of psychopathy is mediated by psychosocial factors, namely the fathers' employment and accommodation problems, and drug use (Auty et al., 2015).

Experiments to Investigate the Double-Edged Sword Theory

For exploring the influence of neurobiological or genetic evidence on judging in criminal cases, several experimental studies with both mock jurors and judges have been performed. All but one of the studies described below are from the United States; only one study comes from Germany (Fuss et al., 2015).

Gurley and Marcus performed the first controlled study to examine the influence of neuroimages and neurological testimony on students' verdicts in non-guilty by reason of insanity cases (Gurley and Marcus, 2008). They found that defendants diagnosed with psychosis were more likely to be judged non-guilty by reason of insanity than those diagnosed with psychopathy. Furthermore, the addition of neuroimages showing brain damage increased the likelihood of such a verdict, as did testimony stating that the defendant's disorder began after a brain injury in a car accident (Gurley and Marcus, 2008).

Greene and Cahill (2012) performed a similar experiment with psychology students acting as mock jurors in a capital case. Consistent with the findings of Gurley and Marcus (2008), they found that both types of neuroscientific evidence had a mitigating effect by reducing the likelihood that jurors would sentence the defendant to death (but only for defendants at high risk of future dangerousness).

Appelbaum and Scurich (2014) investigated the influence of different explanations of impulsivity on the sentencing of jurors that were representative for the United States population. They found that evidence of genetic predisposition for impulsive behavior, including violence, did affect neither whether the defendant was convicted of first- or second-degree manslaughter or first- or second-degree murder, nor the sentence (Appelbaum and Scurich, 2014). However, participants who received evidence of childhood abuse or evidence of childhood abuse plus evidence of genetic predisposition imposed longer sentences (Appelbaum and Scurich, 2014). Genetic evidence and genetic plus childhood abuse evidence engendered the greatest fear of the defendant (Appelbaum and Scurich, 2014).

Recently, Allen et al. (2019) published a modified study design in order to distinguish between different motivations for punishment. They assumed that the question whether a given biological or psychological disorder is treatable has a high impact on juror's decision for the type of custody and for the sentence duration (Allen et al., 2019). They found that both brain evidence and psychological evidence had mitigating effects on prison sentencing, whereby brain evidence had a stronger effect (Allen et al., 2019). However, brain evidence led to decisions for longer involuntary hospitalizations. They found that the variation in

sentencing was explained best by "deontological considerations pertaining to moral culpability" (Allen et al., 2019).

Aspinwall et al. (2012) authors were the first to test experimentally the influence of genetic evidence on sentencing decisions of United States judges. In a nationwide experiment, they presented U.S. state trial judges ($N = 181$) a hypothetical case vignette, which was a modification of the famous case of *Mobley v. State* (Mobley v. The State, 1995; Mobley v. Head, 2001). In the case vignette, the offender was convicted of aggravated battery (instead of murder as in the real case). All participants received a psychiatric testimony about the offender's psychopathy. The study used a 2×2 design. One group was told that the psychiatric testimony was presented by the defense; the other one that it was presented by the prosecution. One group received the explanation that the offender's psychopathy was related to his low-activity MAOA genotype; the other group did not receive any genetic explanation. The judges were randomly assigned to one of these four groups. The authors found that the judges considered the psychiatric testimony about psychopathy aggravating. The additional presentation of neurogenetic evidence for the offender's psychopathy significantly reduced sentencing (from 13.9 to 12.8 years) (Aspinwall et al., 2012).

Fuss et al. (2015) repeated Aspinwall's study in order to investigate whether the double-edged sword effect can also be found in German judges. They found that neurogenetic evidence significantly reduced the German judges' estimation of legal responsibility of the convict. Nevertheless, the average prison sentence was not influenced. Most interestingly, neurogenetic evidence presented by the prosecution significantly increased the number of judges (23% compared with 6%) ordering an involuntary commitment in a forensic psychiatric hospital (Fuss et al., 2015). The different results of these two studies show that the judges' responses to neurogenetic evidence is highly influenced by the legal system in which they operate (Fuss et al., 2015).

The legal theorists Denno and McGivney (2013) have strongly criticized Aspinwall's study as significantly flawed due to problems with both the design and the methodology. Their main points of criticism are: (1) The hypothetical defendant is featured with psychopathy, although this condition is not fully recognized in the medical community and not listed in the current or any prior edition of the DSM. Indeed, the defendant in the real-life case upon which the study's hypothetical case is based claimed that he had an antisocial personality disorder. (2) The study authors instructed the participants that rehabilitation was not an alternative for the offender, because treatment has been ineffective for adult psychopaths so far. This directive substantially loaded the dice in favor of the judges' sentencing decisions being influenced by considerations of future dangerousness or retribution. (3) The study did not include a control group, which was not told that the offender was diagnosed with psychopathy. (4) In contrast to the real-life case, the study's defendant did not commit murder, but only an aggravated assault. Insofar, the study's hypothetical case differs significantly from a typical behavioral genetics criminal case, which involve capital crimes. (5) The study does not describe the gene-environment interaction that is present in nearly any real-world criminal case involving behavioral genetics evidence (Denno and McGivney, 2013).

Denno and McGivney (2013) conclude that Aspinwall's study may interpret the effects of genetics evidence as a double-edged sword, but that there is no support for such a simplistic perspective in actual case law nor are the evidentiary hurdles the same for each side of that sword. It is much more difficult for the State to prove that genetic factors will predict a defendant's future dangerousness than it is for the defense to introduce such information to suggest why a defendant should not be executed (Denno and McGivney, 2013).

Denno and McGivney (2013) emphasize that Denno's (2012) comprehensive survey of criminal cases involving behavioral genetics evidence did not reveal a single case in which such evidence was used to support the likelihood of a defendant's future dangerousness. According to Denno's survey, there was no case in which the State introduced behavioral genetics evidence in any capacity, much less as an aggravating factor. To the contrary, only defense attorneys introduced behavioral genetics evidence into court (Denno and McGivney, 2013).

Objective and Conception of the Present Study

The main objective of this study is to investigate the influence of different types of neurobiological explanations on the sentencing decisions made by German law students. In particular, we wanted to find out whether and to what extent neurobiological explanations influence the students when it comes to evaluating the legal and moral responsibility of a psychopathic offender, deciding about a prison sentence or forensic psychiatric hospital confinement, and to sentencing. Thereby, we compared two different neurobiological explanations (namely, a genetic explanation and a brain injury explanation) with no neurobiological explanation.

The present study is based on the studies of Aspinwall (Aspinwall et al., 2012) and Fuss (Fuss et al., 2015). However, we modified their concept in order to address some of Denno and McGivney's (2013) criticism.

First, we presented a case of manslaughter (as in the real case *Mobley v. State*) instead of aggravated assault (as in the studies of Aspinwall and Fuss), because most real criminal cases, in which genetic evidence is presented, are capital crime cases.

Second, we did not establish two groups of which one was told that the genetic evidence was presented as mitigating by the defense, and the other one was told that it was presented as aggravating by the prosecution. The latter case is unrealistic according to Denno's surveys (Denno, 2012, 2015). Particularly, for Germany, this case is unrealistic.

Third, we established three different groups: the first group received genetic evidence, the second group received neurobiological evidence for a brain trauma, and the third group did not receive any biological evidence.

A further difference is that we interviewed law students instead of judges. The main reason for this decision was that we wanted to achieve a high response rate. We estimate that the response rate among German judges in the study of

Fuss (Fuss et al., 2015) is only about 2%. [Only 375 judges responded, although in 2016, there were more than 15,000 judges at ordinary courts in Germany (Statista, 2018).] Due to the extreme lack of judges and the severe overload of the German courts, we expected that even fewer judges would participate in a new survey among judges. A small response rate is generally associated with a strong bias. In order to receive a high response rate, we decided for investigating law students instead of judges. In our experience, nearly all students participate in surveys, which are recommended by their professors and conducted directly after the courses. A further reason for investigating law students was that they are the future decision-makers in criminal cases, and they are particularly influenced by university professors and thus by recent developments in legal theory.

MATERIALS AND METHODS

Participants

We recruited 317 law students from three major German universities in the summer semester of 2018. We invited the students after the lecture classes to participate in the survey. They did receive neither course credit nor an allowance for participating. We informed the students about the voluntariness of participation, about the study purposes and procedures, which guaranteed full anonymity and compliance with the EU General Data Protection Regulation. Ethics approval by the Local Ethics committee of Charité – Universitätsmedizin Berlin was not applicable given the study design, purpose, and procedures.

Design

This prospective quasi-experimental study used case vignettes as independent variables (Aguinis and Bradley, 2014; Auspurg and Hinz, 2014). We developed a specific case vignette to assess the influence of three different explanations of psychopathy in a criminal case of manslaughter. All three vignettes contain the same information about a forensic expert's testimony that is said to report compelling evidence for the diagnosis of "psychopathy." The independent variable being manipulated is the type of information supporting the expert diagnosis: either no biological explanation of "psychopathy" versus a neurobiological explanation (brain injury) versus a genetic explanation (MAOA gene). All three vignettes contained the same set of instructions and background information based on the German Penal Code. The outcome measure is a questionnaire on legal and moral responsibility, free will, the type of custody, and the duration of the sentence. The vignettes and the questionnaire can be found in the **Supplementary Material**.

Participants from all three universities were allocated to three experimental conditions (**Figure 1**). Each participant received a questionnaire asking for demographic information and presenting one of three types of vignettes. Each vignette initially presented the exact same content and phrasing of a criminal case. The case describes a young man who committed manslaughter of his former girlfriend. All vignettes reported that a psychiatric expert had assessed the perpetrator as a psychopath. Each vignette gave a different etiological explanation for the psychopathy of the perpetrator, depending on the experimental condition. The full text of the vignettes is presented in the **Supplementary Material**. Participants received all textual information in German translated by a German native speaker (S.M.). We collected the data in the form of a paper and pencil questionnaire.

For analytical reproducibility, we openly publish the statistical analyses as R scripts together with the full data set on the Open Science Framework website[1].

Statistical Analysis

We conducted seven separate fixed-effect Analyses of Variance (ANOVA) to examine group differences (main effect). The seven dependent variables are (1) legal responsibility, (2) moral responsibility, (3) free will, (4) the type of custody assigned, (5) the duration of sentencing, (6) the influence of the biological explanation on the type of custody assigned, and (7) the influence of the biological explanation on the duration of sentencing.

As between-subjects factor we used group allocation ("Absent Biomechanism" group versus "Brain Injury" group versus "MAOA Gene" group). We included all demographic variables consisting of gender, number of semesters, level of biology training, the acquaintance with psychopathy and home university as between-subject factors into the analyses. For main effects of group differences, a strict alpha-level of 0.005 was used due to multiple testing of a family of related hypotheses about the influence of the case vignette on the judgment of the law students and the associated risk of an inflated false-positive rate (Benjamin et al., 2018).

Post hoc t-tests were performed in the event of significant group differences according to the conventional alpha-level of 0.05 but were considered "exploratory" if above the predefined alpha-level for the main effects (0.005). With "exploratory," we mean that the effect warrants replication but can be considered suggestive to devise new hypotheses. The multiple *post hoc* pairwise-comparisons can determine between which specific pairs of groups, the difference of the means is statistically significant. Given unequal group sizes, the non-parametric Games–Howell test was chosen over the more common Tukey's *post hoc* test because the former does not make assumptions about normality, equal variances, or sample sizes.

For ANOVA, partial η^2 was used as effect size (small effect $\geq$ 0.01; medium effect $\geq$ 0.06; large effect $\geq$ 0.14). Exploratory χ^2-tests were used to examine potential differences in demographic characteristics between the three groups (**Table 1**).

Missing values arising from incomplete survey responses were less than 3% and imputed using non-parametric random forests (Stekhoven and Bühlmann, 2012). *A priori* Power to detect a

[1]https://osf.io/6r5ng/

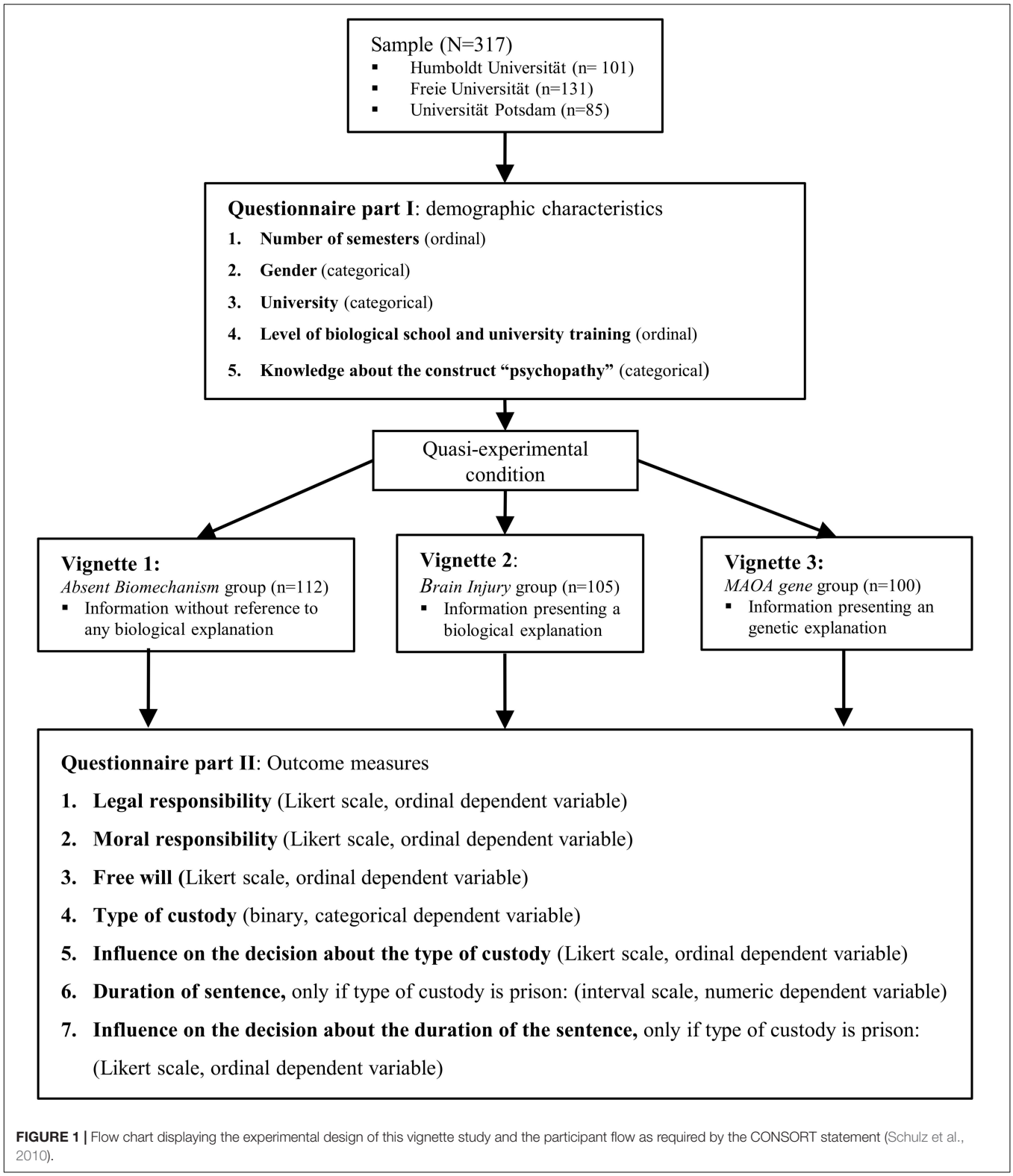


FIGURE 1 | Flow chart displaying the experimental design of this vignette study and the participant flow as required by the CONSORT statement (Schulz et al., 2010).

medium effect size or larger in a balanced three group ANOVA with $\alpha = 0.005$ and a Power = 90% was estimated to be optimal with a total sample size of $n = 318$.

All statistical analyses were carried out in R version 3.5.2.

RESULTS

In total, 317 law students returned the questionnaire at least partially answered. Overall, 1.2% of the questionnaire

TABLE 1 | Comparison of baseline characteristics between the groups.

	Group				
	Absent	**Brain injury**	**MAOA gene**	**Row sum**	
Gender					
Female	57 (18.4%)	55 (17.8%)	60 (19.4%)	172 (55.6%)	$\chi^2(2, N = 309) = 2.082$, *Cramer's V* = 0.082, $p = 0.353$
Male	54 (17.5%)	45 (14.6%)	38 (12.3%)	137 (44.4%)	
Total	111 (35.9%)	100 (32.4%)	98 (31.7%)	309 (100%)	
University					
Humboldt-Universität zu Berlin	34 (10.7%)	44 (13.9%)	23 (7.3%)	101 (31.9%)	$\chi^2(4, N = 317) = 14.76$, *Cramer's V* = 0.153, $p = 0.005$
Freie Universität Berlin	55 (17.4%)	30 (9.5%)	46 (14.5%)	131 (41.4%)	
Universität Potsdam	23 (7.3%)	31 (9.8%)	31 (9.8%)	85 (26.9%)	
Total	112 (35.3%)	105 (33.1%)	100 (31.5%)	317 (100%)	
Level of Biology training					
Grammar school until 10th grade	9 (2.8%)	14 (4.4%)	12 (3.8%)	35 (11%)	χ^2 (4, $N = 317$) = *19.272, Cramer's V* = 0.174, $p = 0.001$
Until university entrance diploma	83 (26.2%)	89 (28.1%)	82 (25.9%)	254 (80.2%)	
University classes (Biology/Medicine)	20 (6.3%)	2 (0.6%)	6 (1.9%)	28 (8.8%)	
Total	112 (35.3%)	105 (33.1%)	100 (31.5%)	317 (100%)	
Semester					
First	2 (0.6%)	3 (0.9%)	0 (0%)	5 (1.5%)	$\chi^2(18, N = 317) = 32.283$, *Cramer's V* = 0.226,
Second	53 (16.7%)	28 (8.8%)	46 (14.5%)	127 (40%)	*Fisher's p* = 0.005
Third	0 (0%)	2 (0.6%)	2 (0.6%)	4 (1.2%)	
Forth	56 (17.7%)	64 (20.2%)	46 (14.5%)	166 (52.4%)	
Fifth	0 (0%)	2 (0.6%)	0 (0%)	2 (0.6%)	
Sixth	1 (0.3%)	3 (0.9%)	3 (0.9%)	7 (2.1%)	
Seventh	0 (0%)	1 (0.3%)	0 (0%)	1 (0.3%)	
Eighth	0 (0%)	0 (0%)	2 (0.6%)	2 (0.6%)	
Eleventh	0 (0%)	0 (0%)	1 (0.3%)	1 (0.3%)	
Twelfth	0 (0%)	2 (0.6%)	0 (0%)	2 (0.6%)	
Total	112 (35.3%)	105 (33.1%)	100 (31.5%)	317 (100%)	
Acquaintance with psychopathy					
Nothing at all	28 (9%)	17 (5.5%)	16 (5.1%)	61 (19.6%)	$\chi^2(12, N = 311) = 20.706$, *Cramer's V* = 0.182,
Movies	3 (1%)	8 (2.6%)	3 (1%)	14 (4.6%)	*Fisher's p* = 0.051
Fictional literature	5 (1.6%)	3 (1%)	2 (0.6%)	10 (3.2%)	
School	8 (2.6%)	6 (1.9%)	5 (1.6%)	19 (6.1%)	
Popular science magazines	6 (1.9%)	5 (1.6%)	1 (0.3%)	12 (3.8%)	
TV documentations	28 (9%)	28 (9%)	17 (5.5%)	73 (23.5%)	
Scientific literature	34 (10.9%)	36 (11.6%)	52 (16.7%)	122 (39.2%)	
Total	112 (36%)	103 (33.1%)	96 (30.9%)	311 (100%)	

response are incomplete. Most participating law students were enrolled at the Freie Universität Berlin (41.4%), followed by Humboldt-Universität zu Berlin (31.9%) and Universität Potsdam (26.9%). 55.6% of the law students were female and 44.4% were male. **Table 1** shows the distribution of the sample characteristics between the experimental groups.

Moral Responsibility

Fixed-Effects ANOVA indicate no significant differences between the groups according to our predefined criterion of significance [$F(2, 294) = 5.15$, $p < 0.006$, $\eta_p^2 = 0.03$]. The partial $\eta^2 = 0.03$ and 90% CI suggest that this effect is of small effect size [0.01, 0.07]. Exploratory *post hoc* *t*-tests revealed no significant differences between the "Absent Biomechanism" group and the "Brain Injury" group [$t(214.79) = 1.80$, $p = 0.171$], nor between the "Brain Injury" group and the "MAOA gene" group [$t(196.16) = 1.08$, $p = 0.527$], or the "MAOA gene" group and the "Absent Biomechanism" group [$t(204.98) = 2.71$, $p = 0.020$] (**Figure 2**).

Free Will

Fixed-Effects ANOVA indicate no significant differences between the groups [$F(2, 294) = 3.95$, $p < 0.02$, $\eta_p^2 = 0.03$]. The partial $\eta^2 = 0.03$ and 90% CI suggest that this effect is of small effect size [0.00, 0.06]. Exploratory *post hoc* *t*-tests revealed no significant differences between the "Absent Biomechanism" group and the "Brain Injury" group [$t(212.85) = 2.42$, $p = 0.043$], nor between the "Brain Injury" group and the "MAOA gene" group [$t(204.94) = 1.17$, $p = 0.475$], or the "MAOA gene" group and the "Absent Biomechanism" group [$t(202.35) = 1.17$, $p = 0.470$] (**Figure 3**).

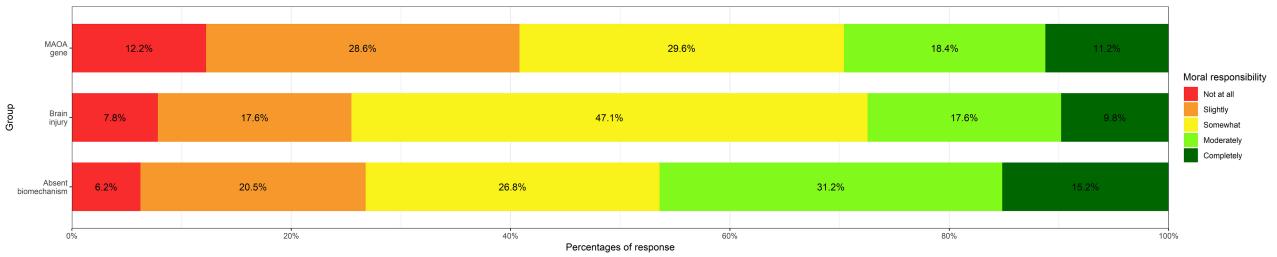


FIGURE 2 | Plot of proportional cross-tables (contingency tables) showing the response to the question how morally responsible the offender is.

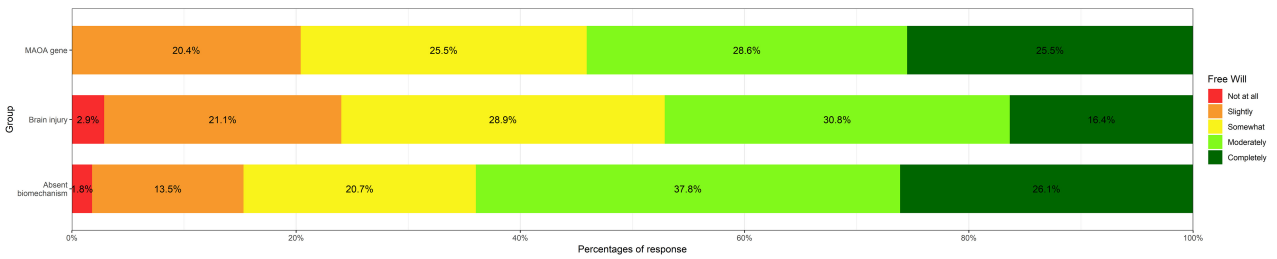


FIGURE 3 | Plot of proportional cross-tables (contingency tables) showing the response to the question to which degree the offender had a free will at the time of manslaughter.

Legal Responsibility

Fixed-Effects ANOVA indicate significant differences between the groups [$F(2, 294) = 8.24$, $p < 0.001$, $\eta_p^2 = 0.05$]. The partial $\eta^2 = 0.05$ and 90% CI suggest that this effect is of small to possibly moderate effect size [0.02, 0.10]. *Post hoc t*-tests revealed significant differences between the "Absent Biomechanism" group and the "Brain Injury" group [$t(213.04) = 3.27$, $p = 0.004$], i.e., the group that received no biological explanation assigned a higher legal responsibility. The mean response in the "Absent Biomechanism" group is 2.29 (SD = 0.53) and in the "Brain Injury" group 2.05 (SD = 0.54) with the response "1" meaning "not at all legally responsible," "2" being "diminished legally responsible," and "3" being "fully legally responsible." Thus, most students answered "diminished legally responsible," but in the "Brain Injury" group significantly more students considered the perpetrator "not at all" legally responsible compared to the "Absent Biomechanism" group. In the latter group, significantly more students responded "fully legally responsible" compared to the "Brain Injury" group (**Figure 4**).

Post hoc t-tests revealed no significant differences between the "MAOA gene" and the "Brain Injury" group [$t(200.53) = 2.20$, $p = 0.074$], nor between the "Absent Biomechanism" group and the "MAOA gene" group [$t(201.43) = 0.86$, $p = 0.666$].

Type of Custody

Fixed-Effects ANOVA indicate no significant differences between the groups [$F(2, 302) = 4.12$, $p < 0.017$, $\eta_p^2 = 0.03$]. The partial $\eta^2 = 0.03$ and 90% CI suggest that this effect is of small effect size [0.00, 0.06]. *Post hoc t*-tests revealed no significant differences between the "Absent Biomechanism" group and the "Brain Injury" group [$t(215.00) = 2.26$, $p = 0.063$], nor between the "Brain Injury" group and the "MAOA gene" group [$t(209.01) = 1.83$, $p = 0.163$], or the "MAOA gene" group and the "Absent Biomechanism" group [$t(202.05) = 0.39$, $p = 0.918$] (**Figure 5**).

Duration of Sentencing

On a descriptive level, the mean prison sentence assigned by law students differed only slightly and the group differences were not significant. In the "Absent Biomechanism" group, the mean prison sentence assigned was 9.15 years (SD = 3.47 years), in the "Brain Injury" group 10.06 years (SD = 5.37 years) and in the "MAOA Gene" group 10.54 years (SD = 3.84 years) (**Table 2**).

Fixed-Effects ANOVA indicate no significant differences between the groups [$F(2, 106) = 0.64$, $p < 0.530$, $\eta_p^2 = 0.01$]. The partial $\eta^2 = 0.01$ and 90% CI suggest that this effect is of zero to very small effect size [0.00, 0.05].

For detailed descriptions of the results for all statistical analyses calculated, see **Supplementary Statistical Analysis**.

One plausible suggestion is that the level of expertise and background knowledge can influence decision-making. For this reason, we included variables such as the number of semesters of the participants, their home university, their level of biological training and their acquaintance with psychopathy as covariates in the fixed effect ANOVA. Doing so allowed us to study the influence of these factors on the outcome such as sentencing. However, in our sample, these factors did not had a significant impact [number of semesters: $F(1, 106) = 0.06$, $p < 0.814$; level of biological training: $F(2, 106) = 0.03$, $p < 0.970$; acquaintance with psychopathy: $F(6, 106) = 0.37$, $p < 0.895$]. We had a very homogeneous sample with 92.4% of the students being in semester 2 or 4, and with 80.2% having their biology knowledge

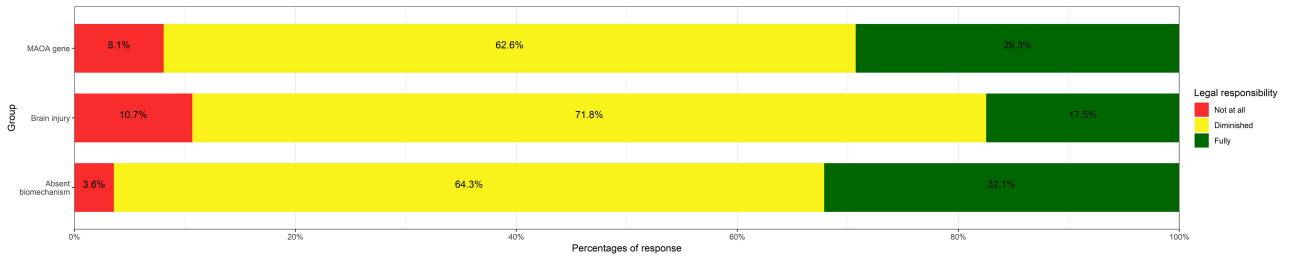


FIGURE 4 | Plot of proportional cross-tables (contingency tables) showing the response to the question how legally responsible the offender is.

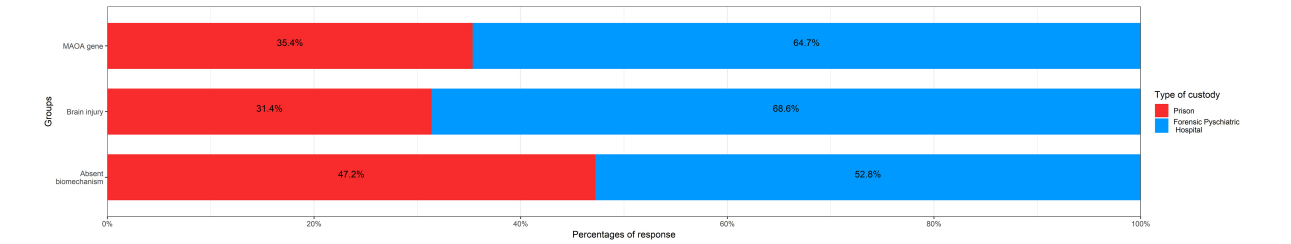


FIGURE 5 | Plot of proportional cross-tables (contingency tables) showing the response to the question, which type of custody should be assigned to the offender depending on the group of respondents.

from school until the university entrance diploma (**Table 1**). We only found an association between the level of biological training and the evaluation of moral responsibility, which we consider exploratory as the p-value is greater than $p = 0.005$ [$F(2, 294) = 4.11$, $p < 0.017$]. This may give rise to the hypothesis that the level of biological training affects the decision-making of law students with regard to the assessment of moral responsibility, but further research directly addressing this hypothesis would be needed to investigate this hypothesis.

Influence of Expert Testimony on Decision-Making

We also examined whether participants noticed being influenced in their decision-making by the expert testimony. As described in further detail in the **Supplementary Materials S6**, **S7**, there was no significant difference between the participants' responses of the three groups to the question whether the expert testimony affected the decision to assign a prison sentence or custody in a forensic hospital [$F(2,302) = 0.38$, $p < 0.685$.]. There was also no group difference with regard to the participants' responses to the question whether the expert testimony affected the duration of prison sentencing assigned [$F(2,105) = 0.69$, $p < 0.505$].

TABLE 2 | Duration of sentencing depending on the group of respondents (M = mean, SD = standard deviation, n = number of students in the group).

Group	M (SD) [years in prison]	*n*
Absent biomechanism	9.15 (3.47)	53
Brain injury	10.06 (5.37)	33
MAOA gene	10.54 (3.84)	35

DISCUSSION

Neuroscientific evidence has been increasingly introduced in criminal trials all around the world to explain criminal behavior. Our results indicate that the different neurobiological information has only small effects on the assessment of law students. However, neurobiological information is often used when very high stakes are involved such as death penalty or the verdict "not guilty" in capital crimes. In such contexts, every bit of information that influences human judgment plays a decisive role. In our study, the strongest effect was observed with regard to legal responsibility. Law students were asked to rate the legal responsibility of a perpetrator after having received one of three different kinds of information about the perpetrator. Overall, there was a significant difference in the assessment of legal responsibility of the law students depending on the kind of information received. Pair-wise comparisons of the groups showed that students who received information describing a major brain injury of the perpetrator rated the legal responsibility significantly lower than the students who did not receive a biological explanation. However, no similar effect was found for information describing a MAOA gene susceptibility for psychopathy.

Our results can be compared to previous findings of the studies of Aspinwall (Aspinwall et al., 2012) and Fuss (Fuss et al., 2015), although we modified the study design in light of Denno and McGivney's (2013) criticism. Our sample size ($N = 317$) is comparable to the above-mentioned studies (Aspinwall et al., 2012: $N = 181$; Fuss et al., 2015: $N = 372$).

Similar to our main finding of reduced legal responsibility in case of a brain injury, the legal responsibility in the study of Fuss (Fuss et al., 2015) was significantly lower in the

group that received biomechanistic information compared to no biological information.

In addition to legal responsibility, we examined the influence of neuroscientific evidence on law students' assessment of moral responsibility, of free will, the type of custody, and of the duration of sentence. Due to the multiplicity of statistical analyses, we used a strict criterion for significance of $p < 0.005$ as recommended by Benjamin et al. (2018) and considered the $p < 0.05$ as exploratory. The effect of brain injury evidence on legal responsibility was the only effect that was significant given the strict criterion. However, on a more exploratory interpretation, we observed some suggestive differences of the influence of neuroscientific evidence on the students' assessment of moral responsibility and free will. The law students who received biological information about the MAOA gene, tended to assign less moral responsibility compared to the students who received no biological explanation. In addition, the free will of the perpetrator was assessed to be lower by the group of students who received information about a brain injury compared to students who received no biological explanation. Due to the exploratory character of these analyses, these findings warrant replication in an independent sample before taken to represent real effects.

The mean prison sentence assigned by law students differed only slightly, and the group differences were not significant ("Absent Biomechanism" group: 9.15 years, "Brain Injury" group: 10.06 years, "MAOA Gene": 10.54 years).

In comparison, in Aspinwall's study (Aspinwall et al., 2012) the mean prison sentence was higher than in our sample. Important to note is that the mean prison sentence was lower in the group with genetic evidence (12.83 years) than in the group without biological explanation (13.93 years) (Aspinwall et al., 2012). In Fuss' study (Fuss et al., 2015), the average prison sentence was not affected by the presentation of neurogenetic evidence. Also in our study, the prison sentence was not influenced by the presentation of biological evidence.

LIMITATIONS AND STRENGTHS

The ecological validity of this study can be doubted, as well as that of all previous studies. As Scurich (2018) has recently pointed out, it is doubtful that a cursory written expert's report is remotely similar to a real expert's testimony who presents his results in court, uses PowerPoint presentations and is subjected to cross examination. For future studies, more realistic simulations should be used to increase the ecological validity.

This is a prospective study with experimental control to increase internal validity. Since law students are the upcoming judicial decision-makers in the legal system, the study is also externally valid in respect of the population examined. However, quasi-experimental study designs come with certain risk of bias. In particular, the lack of randomization prevents any strong claims ruling out that non-measured variables confound the results. For mitigating the risk of bias, we included theoretically relevant demographic differences between groups as between-factors into the statistical analysis. *Post hoc* tests were adjusted for multiple comparison and the statistical power was sufficient to find at least medium effects.

A strength of our paper is that we rigorously corrected the significance level for multiple testing (Benjamin et al., 2018).

A further strength is that our statistical analysis accounts for the influence of demographic factors such as gender, the number of semesters, and level of biological education.

CONCLUSION

The main purpose of this research was to understand the judgments of German law students depending on two different types of neuroscientific evidence being presented in the courtroom. The question is whether biological information influences the judgment of law students. Indeed, the "Brain Injury" group evaluated the perpetrator less legally responsible than the "Absent Biomechanism" group.

The question whether neuroscientific and genetic evidence in criminal cases is a double-edged sword or not, has been answered differently by different studies. The results from the experimental studies from Germany (Fuss et al., 2015, and the present study) are partly inconsistent with the results from the experimental studies from the United States (Gurley and Marcus, 2008; Aspinwall et al., 2012; Greene and Cahill, 2012; Appelbaum and Scurich, 2014; Allen et al., 2019). In contrast to the United States studies, the German studies did not find a mitigating effect of neuroscientific evidence in terms of the duration of sentencing. The study of Fuss found that neurogenetics evidence leads to more decisions for forensic psychiatric hospital, which has the consequence of a longer and indefinite detention (Fuss et al., 2015).

For investigating whether a double-edged sword exists, it is important to compare the results of surveys of real criminal cases from different countries, too. Indeed, they provide mixed results, too.

For the United States, Denno has comprehensively analyzed criminal cases from the United States, of which 553 addressed neuroscience evidence for the defendant (from 1992 to 2002) (Denno, 2012), and 81 addressed behavioral genetics evidence (including family history evidence and MAOA deficiency evidence) (from 1994 to 2007) (Denno, 2015). Denno's studies systematically investigated how United States courts assess the mitigating and aggravating effects of neuroscience or genetic evidence, respectively. She found that neuroscience evidence is typically raised in cases where defendants are facing the death penalty, a life sentence or a substantial prison sentence (Denno, 2015). Usually neuroscience evidence is offered to mitigate punishments in the way that traditional criminal law has always allowed, especially in the penalty phase of death penalty trials (Denno, 2015). Neuroscience evidence is only rarely used to bolster a defendant's future dangerousness (Denno, 2015). In the rare cases when prosecutors utilized neuroscientific evidence to implicate a defendant's propensity to commit crimes, they typically did so only by building upon the evidence first introduced by a defense expert (Denno, 2015). The same is valid for behavioral genetics evidence, which has been applied

almost exclusively as mitigating evidence in death penalty cases (Denno, 2012). Between 2007 and 2011, the State never presented behavioral genetics evidence as aggravating evidence or for indicating the future dangerousness of the defendant (Denno, 2012). United States courts accept both neuroscience and behavioral genetics evidence (Denno, 2012, 2015). They even expect attorneys to raise neuroscience evidence when possible on behalf of their clients. Courts grant defendants their "ineffective assistance of counsel" claims when attorneys fail to pursue mitigating neuroscience or genetic evidence (Denno, 2015). Sometimes courts even penalize attorneys who neglect the obligation to pursue mitigating neuroscience evidence (Denno, 2015). Denno concludes that her study "controverts the popular image of neuroscience evidence as a double-edged sword – one that will either get defendants off the hook altogether or unfairly brand them as posing a future danger to society" (Denno, 2015).

Farahany (2015) also examined the use of neurological and behavioral genetic evidence in United States criminal law. For that, she and her team investigated 1,585 judicial opinions issued between 2005 and 2012. Although many scientists discredit the use of neurobiological evidence in criminal law, and some call for "an outright ban on its use" due to significant methodological problems, Farahany (2015) concludes that neuroscience is "already entrenched in the United States legal system." She found that neurobiological evidence is increasingly used in criminal cases (Farahany, 2015). Neurobiological evidence is used broadly, and is not limited to capital cases as mitigating evidence (Farahany, 2015). Farahany states that neurobiological evidence is "in a rarified position of must-investigate evidence," and summarizes: "Defense counsels are ineffective if they fail to mount a defense at all, sleep through an entire (but not just parts of) a trial, or if they fail to investigate a probable neurobiological abnormality in a defendant." (Farahany, 2015).

In the Netherlands, neuroscientific and genetic evidence is in most cases no double-edged sword. According to De Kogel and Westgeest's (2015) analysis of 231 criminal cases published between 2000 and 2012, neuroscientific evidence is introduced as mitigating evidence in the majority of the cases found. Only in some cases, defendants were considered diminished or not responsible for their crime, but received a longer sentence, such as a custody in a forensic psychiatric hospital that can be periodically extended. In some other cases, the defendants did not receive longer sentences despite their "untreatable" neurobiological deficits, when the experts saw room for reduction of recidivism risk (De Kogel and Westgeest, 2015).

For England and Wales, Catley and Claydon (2015) analyzed 204 criminal cases from 2005 to 2012. They found that most appellants, who used neuroscientific evidence when they appealed against conviction, were unsuccessful. However, in the few successful cases, the neuroscientific evidence had nearly always a central role in the successful appeal (Catley and Claydon, 2015). The authors do not discuss whether they found the double-edged sword effect in the cases analyzed.

However, in Canada, neuroscientific evidence is a double-edged sword for criminal offenders according to Chandler's (2015) analysis of 133 criminal cases published between 2008 and 2012. In Canada, the most common form of biological evidence considered is fetal alcohol spectrum disorder, followed by medical history of traumatic brain injury and neuropsychological testing (Chandler, 2015). Functional MRI investigations and genetic tests did not play any role in the court decisions analyzed (Chandler, 2015). Chandler (2015) found that neuroscientific evidence suggested diminished capacity, but also tends to increase judgements about risk and dangerousness given the view that brain injuries can sometimes be managed but not cured.

For Australia, Alimardani and Chin (2018) found on grounds of a systematic review that in some cases, neuroscientific evidence presents a double-edged sword. It can serve to either aggravate or mitigate a sentence. Because the courts also consider the protection of society, a sentence can be prolonged when neuroscientific evidence suggests that the offender poses a particular risk of re-offending. On the other side, neuroscientific evidence can suggest a reduced risk of future offending, and thus support a more lenient sentence. Furthermore, neuroscientific evidence can mitigate the offender's moral culpability and thus reduce the significance of general deterrence, so that the sentence can be mitigated. In most cases analyzed by the authors, neuroscience evidence only leads to mitigation and was rarely used as evidence for the offender's risk of recidivism (Alimardani and Chin, 2018).

For Iran, Alimardani (2018) has investigated the potential applicability of neuroscientific evidence in the criminal justice system. He demonstrates that neuroscientific evidence can be used *inter alia* both for establishing the insanity defense and for mitigating the sentence for some kinds of crimes. He concludes that neuroscientific evidence can result on the one hand in a successful defense of insanity and thus in the offender's discharge. On the other hand, if it indicates a condition, which can put the society in danger, the offender will be detained in a psychotherapeutic facility for an indeterminate period (Alimardani, 2018). Therefore, neuroscientific evidence may be a double-edged sword in the Iranian criminal justice system.

For Germany, an investigation of the influence of neuroscientific and genetic evidence in real criminal court cases has not been published so far to the best of our knowledge.

In summary, it can be said that neuroscience evidence can present a double-edged sword in Canada, Netherlands, Australia, and Iran, but not in the United States. However, even in the countries, in which a double-edged sword effect might occur, neuroscience evidence seems to lead more often to mitigation than to aggravation of sentencing.

Therefore, the question whether neuroscientific and genetic evidence in criminal cases is a double-edged sword, cannot be answered in general. Rather, the answer depends strongly on the given system of criminal justice. In the United States, punishment is much harder than in most other Western countries; particularly, other Western countries have abolished the capital punishment since many years. On the other side, in the United States, the standards for admitting mitigating evidence at sentencing are purposefully lax (Segal, 2016). The laxity in the admission of mitigating evidence could be the other side of the coin of extreme severity in sentencing.

Furthermore, the type of criminal justice system (common law system vs. civil law system) presumably has a significant impact on the influence of biological evidence on sentencing. Because the roles of professional and lay judges differ in the different systems, and because neuroscientific evidence influences the two groups of judges presumably in different ways.

In the common law system, the jury of lay judges decides whether to convict or to acquit the defendant, whereas the professional judge decides about the penalty (whether detention should take place in prison or in a forensic hospital and the length of the sentence). In the United States, in criminal law cases, in which the death penalty is a prospective sentence, a "death-qualified jury" has to be established. Such a jury has to be composed of jurors who will fairly consider all punishment options, including the death penalty and life imprisonment.

In the civil law system, a jury is called only in court cases involving serious criminal offenses. The jury's influence on sentencing is much smaller than in the common law system. In Germany, in the case of homicide, the jury court consists of three professional and two lay judges. The lay judges are equal judges, who have a full say in the decision-making process about the guilt of the accused and subsequently on the sentence.

Therefore, it is necessary to carefully distinguish the results of studies, which have investigated the effect of biological evidence on the sentencing of professional judges vs. the sentencing of potential lay judges.

The improvement of the ecological validity of experimental research in this field should be in the focus of future research. For that, it is important that the experimental research learns from the research on real criminal cases and vice versa.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the manuscript/**Supplementary Files**.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DG: data collection. SM: study design. DG and MB: statistical analysis. SM, MB, and DG: writing of the report. SE: scientific advice. SM: revision of the manuscript and response to the reviewers.

ACKNOWLEDGMENTS

We thank the Neurasmus Program and Neurocure Scholarship for their academic and financial support for DG for 2 years. Furthermore, we thank ERA-NET NEURON and the Federal Ministry of Education and Research (BMBF) of Germany for funding the work of SM and MB (grant number: 01GP1621A). We acknowledge support from the German Research Foundation (DFG) and Open Access Publication Fund of Charité – Universitätsmedizin Berlin. We thank all Law professors from Humboldt-Universität zu Berlin, the Freie Universität Berlin, and the Universität Potsdam for having invited their students to participate in the study, and all students who participated. We thank Lucia Reuter for assistance with the translation of the questionnaires to English. We thank the reviewers for their helpful suggestions to improve the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02343/full#supplementary-material

REFERENCES

Aguinis, H., and Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organ. Res. Methods* 17, 351–371. doi: 10.1177/1094428114547952

Alimardani, A. (2018). Neuroscience, criminal responsibility and sentencing in an islamic country: Iran. *J. Law Biosci.* 5, 724–742. doi: 10.1093/jlb/lsy024

Alimardani, A., and Chin, J. (2018). *Neurolaw in Australia: The Use of Neuroscience in Australian Criminal Proceedings*. Dordrecht: Springer Netherlands.

Allen, C. H., Vold, K., Felson, G., Blumenthal-Barby, J. S., and Aharoni, E. (2019). Reconciling the opposing effects of neurobiological evidence on criminal sentencing judgments. *PLoS One* 14:e0210584. doi: 10.1371/journal.pone.0210584

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, Fifth edition (DSM-5®)*. Arlington, VA: American Psychiatric Association.

Appelbaum, P. S., and Scurich, N. (2014). Impact of behavioral genetic evidence on the adjudication of criminal behavior. *J. Am. Acad. Psychiatry Law* 42, 91–100.

Aspinwall, L. G., Brown, T. R., and Tabery, J. (2012). The double-edged sword: does biomechanism increase or decrease judges' sentencing of psychopaths? *Science* 337, 846–849. doi: 10.1126/science.1219569

Auspurg, K., and Hinz, T. (2014). *Factorial Survey Experiments*. Thousand Oaks, CA: Sage Publications, 175.

Auty, K. M., Farrington, D. P., and Coid, J. W. (2015). Intergenerational transmission of psychopathy and mediation via psychosocial risks. *Br. J. Psychiatry* 206, 26–31. doi: 10.1192/bjp.bp.114.151050

Babiak, P., and Hare, R. D. (2006). *Snakes in Suits*. New York, NY: Regan.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., et al. (2018). Redefine statistical significance. *Nat. Hum. Behav.* 2, 6–10. doi: 10.1038/s41562-017-0189-z

Blair, R. J. R. (2013). Psychopathy: cognitive and neural dysfunction. *Dialogues Clin. Neurosci.* 15, 181–190. doi: 10.1016/j.neubiorev.2016.10.022

Brunner, H. G., Nelen, M., Breakefield, X. O., Ropers, H. H., and van Ost, B. A. (1993). Abnormal behavior associated with a point mutation in the structural gene for monoamine oxidase A. *Science* 262, 578–580. doi: 10.1126/science.8211186

Caspi, A., McClay, J., Moffitt, T. E., Mill, J., Martin, J., Craig, I. W., et al. (2002). Role of genotype in the cycle of violence in maltreated children. *Science* 297, 851–854. doi: 10.1126/science.1072290

Catley, P., and Claydon, L. (2015). The use of neuroscientific evidence in the courtroom by those accused of criminal offenses in England and Wales. *J. Law Biosci.* 2, 510–549. doi: 10.1093/jlb/lsv025

Chandler, J. A. (2015). The use of neuroscientific evidence in Canadian criminal proceedings. *J. Law Biosci.* 2, 550–579. doi: 10.1093/jlb/lsv026

Cima, M., Tonnaer, F., and Hauser, M. D. (2010). Psychopaths know right from wrong but don't care. *Soc. Cogn. Affect. Neurosci.* 5, 59–67. doi: 10.1093/scan/nsp051

Crego, C., and Widiger, T. A. (2015). Psychopathy and the DSM. *J. Pers.* 83, 665–677. doi: 10.1111/jopy.12115

Darby, R. R., Horn, A., Cushman, F., and Fox, M. D. (2018). Lesion network localization of criminal behavior. *Proc. Natl. Acad. Sci. U.S.A.* 115, 601–606. doi: 10.1073/pnas.1706587115

De Kogel, C. H., and Westgeest, E. J. M. C. (2015). Neuroscientific and behavioral genetic information in criminal cases in the Netherlands. *J. Law Biosci.* 2, 580–605. doi: 10.1093/jlb/lsv024

Denno, D. W. (2012). Courts' increasing consideration of behavioral genetics evidence in criminal cases: results of a longitudinal study. *Mich. State Law Rev.* 2011, 967–1047.

Denno, D. W. (2015). The myth of the double-edged sword: an empirical study of neuroscience evidence in criminal cases. *Boston Coll. Law Rev.* 56, 493–551.

Denno, D. W., and McGivney, A. A. (2013). What real-world criminal cases tell us about genetic evidence. *Hastings Law J.* 64, 1591–1681.

Dugatkin, L. A. (1992). The evolution of the "con artist". *Ethol. Sociobiol.* 13, 3–18. doi: 10.1016/0162-3095(92)90003-M

Farahany, N. A. (2015). Neuroscience and behavioral genetics in US criminal law: an empirical analysis. *J. Law Biosci.* 2, 485–509. doi: 10.1093/jlb/lsv059

Ferguson, C. J. (2010). Genetic contributions to antisocial personality and behavior: a meta-analytic review from an evolutionary perspective. *J. Soc. Psychol.* 150, 160–180. doi: 10.1080/00224540903366503

Fuss, J. (2016). Legal responses to neuroscience. *J. Psychiatry Neurosci.* 41, 363–365. doi: 10.1503/jpn.160147

Fuss, J., Dressing, H., and Briken, P. (2015). Neurogenetic evidence in the courtroom: a randomised controlled trial with German judges. *J. Med. Genet.* 52, 730–737. doi: 10.1136/jmedgenet-2015-103284

Gao, Y., and Raine, A. (2010). Successful and unsuccessful psychopaths: a neurobiological model. *Behav. Sci. Law* 28, 194–210. doi: 10.1002/bsl.924

Greene, E., and Cahill, B. S. (2012). Effects of neuroimaging evidence on mock juror decision making. *Behav. Sci. Law* 30, 280–296. doi: 10.1002/bsl.1993

Gurley, J. R., and Marcus, D. K. (2008). The effects of neuroimaging and brain injury on insanity defenses. *Behav. Sci. Law* 26, 85–97. doi: 10.1002/bsl.797

Hare, R. D. (1996). Psychopathy: a clinical construct whose time has come. *Crim. Justice Behav.* 23, 25–54. doi: 10.1177/0093854896023001004

Hare, R. D. (2003). *The Hare Psychopathy Checklist-Revised, 2nd ed. (PCL-R)*. Toronto, ON: Multi-Health Systems.

Kean, S. (2014). *Phineas Gage, Neuroscience's Most Famous Patient*. Available at: https://slate.com/technology/2014/05/phineas-gage-neuroscience-case-true-story-of-famous-frontal-lobe-patient-is-better-than-textbook-accounts.html (accessed August 12, 2019).

Kim-Cohen, J., Caspi, A., Taylor, A., Williams, B., Newcombe, R., Craig, I. W., et al. (2006). MAOA, maltreatment, and gene–environment interaction predicting children's mental health: new evidence and a meta-analysis. *Mol. Psychiatry* 11, 903–913. doi: 10.1038/sj.mp.4001851

Lorber, M. F. (2004). Psychophysiology of aggression, psychopathy, and conduct problems: a meta-analysis. *Psychol. Bull.* 130, 531–552. doi: 10.1037/0033-2909.130.4.531

Mealey, L. (1995). The sociobiology of sociopathy: an integrated evolutionary model. *Behav. Brain Sci.* 18, 523–599. doi: 10.1017/S0140525X00039595

Mendez, M. F. (2009). The neurobiology of moral behavior: review and neuropsychiatric implications. *CNS Spectr.* 14, 608–620. doi: 10.1017/s1092852900023853

Mobley v. Head, (2001). *United States District Court for the Northern District of Georgia*.

Mobley v. The State, (1995). *Supreme Court of Georgia*.

Palumbo, S., Marioti, V., Iofrida, C., and Pellegrini, S. (2018). Genes and aggressive behavior: epigenetic mechanism underlying individual susceptibility to aversive environments. *Front. Behav. Neurosci.* 12:117. doi: 10.3389/fnbeh.2018.00117

Pardini, D. A., Raine, A., Erickson, K., and Loeber, R. (2014). Lower amygdala volume in men is associated with childhood aggression, early psychopathic traits, and future violence. *Biol. Psychiatry* 75, 73–80. doi: 10.1016/j.biopsych.2013.04.003

Raine, A. (2002). Biosocial studies of antisocial and violent behavior in children and adults: a review. *J. Abnorm. Child Psychol.* 30, 311–326.

Reidy, D. E., Kearns, M. C., DeGue, S., Lilienfeld, S. O., Massetti, G., and Kiehl, K. A. (2015). Why psychopathy matters: implications for public health and violence prevention. *Aggress. Violent Behav.* 24, 214–225. doi: 10.1016/j.avb.2015.05.018

Reimer, M. (2008). Psychopathy without (the language of) disorder. *Neuroethics* 1, 185–198. doi: 10.1007/s12152-008-9017-5

Rosell, D. R., and Siever, L. J. (2015). The neurobiology of aggression and violence. *CNS Spectr.* 20, 254–279. doi: 10.1017/S109285291500019X

Schulz, K. F., Altman, D. G., and Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Med.* 8:18. doi: 10.1186/1741-7015-8-18

Scurich, N. (2018). What do experimental simulations tell us about the effect of neuro/genetic evidence on jurors? *J. Law Biosci.* 5, 204–207. doi: 10.1093/jlb/lsy006

Segal, J. B. (2016). Inherited proclivity: when should neurogenetics mitigate moral culpability for purposes of sentencing? *J. Law Biosci.* 3, 227–237. doi: 10.1093/jlb/lsw005

Shi, Z., Bureau, J.-F., Easterbrooks, M. A., and Lyons-Ruth, K. (2012). Childhood maltreatment and prospectively observed quality of early care as predictors of antisocial personality disorder features. *Infant Ment. Health J.* 33, 55–69. doi: 10.1002/imhj.20295

Sopromazde, S., and Tsiskaridze, A. (2018). Violent behavior. *Front. Neurol. Neurosci.* 42:106–121. doi: 10.1159/000475696

Statista (2018). *Anzahl der Richter in Deutschland nach Gerichtsart am 31. Dezember 2016*. Available at: https://de.statista.com/statistik/daten/studie/37315/umfrage/anzahl-der-richter-in-deutschland-nach-gerichtsart/ (accessed August 16, 2019).

Stekhoven, D. J., and Bühlmann, P. (2012). MissForest-Non-Parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118. doi: 10.1093/bioinformatics/btr597

Tiihonen, J., Rautiainen, M.-R., Ollila, H. M., Repo-Tiihonen, E., Virkkunen, M., Palotie, A., et al. (2015). Genetic background of extreme violent behavior. *Mol. Psychiatry* 20, 786–792. doi: 10.1038/mp.2014.130

Vassos, E., Collier, D. A., and Fazel, S. (2014). Systematic meta-analyses and field synopsis of genetic association studies of violence and aggression. *Mol. Psychiatry* 19, 471–477. doi: 10.1038/mp.2013.31

Yang, Y., and Raine, A. (2009). Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: a meta-analysis. *Psychiatry Res.* 174, 81–88. doi: 10.1016/j.pscychresns.2009.03.012

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

REVIEW
published: 17 March 2020
doi: 10.3389/fpsyg.2020.00220

Neuroprediction and A.I. in Forensic Psychiatry and Criminal Justice: A Neurolaw Perspective

Leda Tortora[1], Gerben Meynen[2,3], Johannes Bijlsma[2], Enrico Tronci[4] and Stefano Ferracuti[1]*

[1] Department of Human Neuroscience, Sapienza University of Rome, Rome, Italy, [2] Willem Pompe Institute for Criminal Law and Criminology/Utrecht Centre for Accountability and Liability Law (UCALL), Utrecht University, Utrecht, Netherlands, [3] Faculty of Humanities, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [4] Department of Computer Science, Sapienza University of Rome, Rome, Italy

Advances in the use of neuroimaging in combination with A.I., and specifically the use of machine learning techniques, have led to the development of brain-reading technologies which, in the nearby future, could have many applications, such as lie detection, neuromarketing or brain-computer interfaces. Some of these could, in principle, also be used in forensic psychiatry. The application of these methods in forensic psychiatry could, for instance, be helpful to increase the accuracy of risk assessment and to identify possible interventions. This technique could be referred to as 'A.I. neuroprediction,' and involves identifying potential neurocognitive markers for the prediction of recidivism. However, the future implications of this technique and the role of neuroscience and A.I. in violence risk assessment remain to be established. In this paper, we review and analyze the literature concerning the use of brain-reading A.I. for neuroprediction of violence and rearrest to identify possibilities and challenges in the future use of these techniques in the fields of forensic psychiatry and criminal justice, considering legal implications and ethical issues. The analysis suggests that additional research is required on A.I. neuroprediction techniques, and there is still a great need to understand how they can be implemented in risk assessment in the field of forensic psychiatry. Besides the alluring potential of A.I. neuroprediction, we argue that its use in criminal justice and forensic psychiatry should be subjected to thorough harms/benefits analyses not only when these technologies will be fully available, but also while they are being researched and developed.

Keywords: neuroprediction, artificial intelligence, recidivism, forensic psychiatry, risk assessment, neurolaw

OPEN ACCESS

Edited by:
José M. Muñoz, Universidad Europea de Valencia, Spain

Reviewed by:
Vaughn R. Steele, Yale University, United States
Cristina Scarpazza, University of Padova, Italy

****Correspondence:***
Leda Tortora
leda.tortora@hotmail.it

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *24 August 2019*
Accepted: *31 January 2020*
Published: *17 March 2020*

Citation:
Tortora L, Meynen G, Bijlsma J, Tronci E and Ferracuti S (2020) Neuroprediction and A.I. in Forensic Psychiatry and Criminal Justice: A Neurolaw Perspective. Front. Psychol. 11:220. doi: 10.3389/fpsyg.2020.00220

INTRODUCTION

Risk assessment is a crucial component of the criminal justice system. In recent years, there has been a growing interest in the development of new tools and techniques to improve risk assessment in the field of forensic psychiatry and criminal justice (Monahan and Skeem, 2015). Currently, more than 200 violence risk assessment tools, often integrated clinical-actuarial instruments, have been developed to predict violent, antisocial, and sexual behavior (Singh et al., 2014), and their use seems to be vastly increasing in criminal justice settings (Conroy and Murrie, 2007).

The central aim of these methods is to identify high-risk and low-risk offenders correctly. Depending on the jurisdiction, they are used to inform a range of medico-legal decisions, for instance regarding sentencing, parole, civil commitment, death penalty, disposition in juvenile courts, and discharge following findings of insanity (Conroy and Murrie, 2007). In recent years, A.I. (Artificial Intelligence) is being used to enhance the predictive accuracy of risk assessment.

The use of algorithmic risk assessment has grown along with the research in the field of neuroimaging, leading to the development of 'brain-reading' techniques that are, to some limited extent, able to decode mental states based on a person's brain activity (Haynes and Rees, 2006), or to classify people in groups based on their brain structure and functionality (Koutsouleris et al., 2012). A possible forensic application of the technique is to identify dangerous offenders. The combination of A.I. and neuroimaging has led to the development of what can be called 'A.I. neuroprediction,' which is the use of structural or functional brain parameters coupled with machine learning methods to make clinical or behavioral predictions. Perhaps, in the near future, A.I. neuroprediction could be more generally used to predict the risk of recidivism in forensic psychiatry and criminal justice. However, application of such techniques raises legal and ethical issues.

The purpose of this paper is to identify possibilities and challenges regarding the possible future use of A.I. neuroprediction of violence and recidivism in the fields of forensic psychiatry and criminal justice, discussing legal implications and ethical issues. In the next section, we will discuss risk-assessment techniques. In the third section, we consider current 'brain-reading' techniques that use neuroimaging coupled with A.I. In the fourth section, we provide an overview of recent neuroprediction studies using neuroimaging data coupled with A.I. to predict recidivism. In the fifth section, we discuss technological limitations and pitfalls of predictive analysis. Finally, in the sixth section, we discuss the ethical and legal issues raised by the application of these techniques.

RISK ASSESSMENT: THE STATE OF THE ART

In the past two decades, in both the US and Europe, interest in and research on violence risk assessment tools have significantly increased, providing different approaches varying from strictly actuarial tools, based on regression, to algorithmic risk assessment, providing a probabilistic estimate of reoffending, to structured professional judgment (Hart, 1998; Douglas and Kropp, 2002). Initially, actuarial methods dominated the field, but their predictive value remained quite limited, if not disappointing (Fazel et al., 2012).

Risk variables associated with an increased likelihood of an individual acting violently or aggressively include criminogenic needs (individual characteristics that increase the risk of recidivism), demographics, socioeconomic status, and intelligence (Gendreau et al., 1996). Risk factors are typically divided into static factors, that are historical and do not change (e.g., criminal history, offense types, childhood abuse) and dynamic factors that are, in principle, changeable and therefore they provide the opportunity for intervention, modifying future risk (e.g., impulsivity, drug use, social support, job, compliance with treatment). Some dynamic factors are quite stable, while others are more "fluid." Dynamic factors need to be measured multiple times, sometimes within short intervals.

At present, the results of risk assessment tools, however, are far from perfect, especially for long term prediction; current criminal risk assessment tools show poor to moderate accuracy, and a good balance between false positives and false negatives is an issue that should be considered, depending both on the social and political context and on the stage of the criminal justice process in which the tool is used (Douglas et al., 2017). Generally, when a risk assessment tool classifies an individual as low-risk, it is often correct. However, if the tool classifies someone as high risk, this is quite often incorrect, and almost more than half of individuals targeted as high-risk are incorrectly classified (Fazel et al., 2012). False positives (defendants are predicted to re-offend, but they do not) seem to be more common than false negatives (defendants are predicted not to re-offend, but they do) (Fazel et al., 2012).

The result is that many people may be or remain incarcerated, while they do not pose a danger to society. As Fazel et al. (2012) wrote: "One implication of these findings is that, even after 30 years of development, the view that violence, sexual, or criminal risk can be predicted in most cases is not evidence-based." This diagnosis of the current state of affairs makes it important to look for ways to improve risk assessment in forensic psychiatry and criminal justice.

Algorithms hold the promise of performing more accurate predictions of criminal behavior than classic approaches, commonly derived from various forms of regression analyses (Berk and Hyatt, 2015). They can be used to provide measures of individualized risk for future violence and help to make decisions about prevention and treatment, in order to minimize risk factors and accentuating protective ones. Risk assessment tools that incorporate machine learning are already in use in pretrial risk evaluation, sentencing, and rehabilitation (Kehl et al., 2017), and are potentially very useful in judicial decision-making, to guide "decisions regarding bail, probation/parole, court-ordered treatment, and civil commitment" (Poldrack et al., 2018).

A.I. AND NEUROIMAGING

Rapid advances in brain imaging and the growing influence of A.I. technologies in many areas of society, from social networks to health care and police force policies (Berk et al., 2018), have led to interest in the potential use of brain imaging combined with A.I. to improve risk assessment and prediction of future violent behavior.

Over the past decade, there has been a significant development of non-invasive anatomical and functional neuroimaging technologies, yielding a lot of data, and statistical machine learning methods are instrumental for analyzing vast amounts of neural data with increasing precision (Lemm et al., 2011) and modeling high-dimensional datasets (Abraham et al., 2014).

Applying statistical machine learning methods to neuroimaging data is referred to as multi-voxel pattern analysis (MVPA) (Ombao et al., 2017, pp164–169). These methods, unlike conventional univariate approaches that analyze only one location at a time, allow for the identification of spatial and temporal patterns in the data, differentiating between cognitive tasks or subject groups with higher sensitivity, jointly analyzing data from individual voxels within a region (Haynes and Rees, 2006).

Since the advent of MVPA methods, they have become a popular approach in the "neuroimaging of healthy and clinical populations; studies have shown that information present in neuroimaging data can be used to decode" – to some extent – "intentions and perceptual states, as well as discriminate between healthy and diseased brains" (Bray et al., 2009). MVPA has been applied to decode visual features like edge orientation (Kamitani and Tong, 2005), the intention to perform one task rather than another (Haynes et al., 2007), sequential stages of task preparation (Bode and Haynes, 2009), and lie detection (Davatzikos et al., 2005; Blitz, 2017, pp. 45–58). While conventional functional imaging studies compare brain activity during different experimental conditions to identify which brain regions are activated by particular tasks, application of MVPA for brain-reading uses "patterns of brain activity to perform a reverse inference and decide what subjects are looking at or thinking about" (Cox and Savoy, 2003; Bray et al., 2009).

These techniques can be considered 'brain-reading' or 'mind-reading' techniques; they combine statistical machine-learning methods with neuroimaging data to reveal information about the brain/mind. Brain-reading has often been studied in the domain of visual perception, where it aims to show how experiences are encoded in the brain. Researchers recently succeeded in training a deep neural network[1] to perform visual image reconstruction from the brain (Shen et al., 2019), decode visual content of dreams (Horikawa et al., 2013), and decode what the brain is 'seeing' by using A.I. to analyse fMRI scans from subjects watching videos (Wen et al., 2017). Despite promising findings, these methods still show many limitations that make it unlikely that a 'general mind-reading technique' will appear in the very near future. Nonetheless, the first simple applications have begun to emerge, including brain-computer-interfaces, studies on lie-detection and approaches for prediction of consumer decisions in the field of neuromarketing (Haynes, 2012, pp. 29–40).

Apart from making inferences regarding the occurrence and nature of mental states (Haynes, 2012, pp. 29–40), another field of application of MVPA techniques is classification. For example, it has been reported that it is possible to predict disease onset by distinguishing individuals within a group based on brain activity or classifying individual people into groups based on the brain data identifying patterns of brain activity or structures (Koutsouleris et al., 2012). Treatment responders can be distinguished from non-responders, by extracting patterns of activity or structural abnormalities that are predictive of abnormal cognitive development and particularly relevant for prediction of clinical outcomes from neuroimaging data (Bray et al., 2009). Some models are applied to discriminate between clinical groups such as Alzheimer Disease patients and cognitively normal elderly individuals (Klöppel et al., 2008), Parkinson's disease patients and healthy controls (Rubbert et al., 2019), schizophrenic patients and healthy controls (Kim et al., 2016), or to detect brain function disorders, such as Autism and attention deficit hyperactivity disorder (ADHD) (Heinsfeld et al., 2018; Sen et al., 2018) and to discriminate between levels of personality traits, for example psychopathy (Steele et al., 2015).

Interesting results have also been reported about prediction of addiction outcomes; machine learning classifiers were able to predict substance abuse treatment completion in a prison inmate population using event-related potentials (ERPs) (Steele et al., 2014; Fink et al., 2016) and functional network connectivity (FNC) analyses of fMRI data (Steele et al., 2018). Furthermore, it turned out to be possible to identify 'neural fingerprints' to predict cocaine abstinence during treatment using CPM, a recently developed machine learning approach (Yip et al., 2019).

A.I. NEUROPREDICTION OF RECIDIVISM

Behavioral traits can be correlated, sometimes strongly, with features of the human brain, and this raises new possibilities for predictive algorithms to be developed, allowing the prediction of dispositions of an individual. These methods are referred to as "neuroprediction," that is the use of structural or functional brain variables to predict prognoses, treatment outcomes, and behavioral forecasts (Morse, 2015). Even though at present it may sound like science fiction, with the continuing development of non-invasive neuroimaging techniques coupled with the growth in the computational power of algorithms, A.I. neuroprediction of recidivism is likely to become available in the near future.

Although there is still need to collect biomarkers of the "criminal" brain, research in the field of neurocriminology has generally focused on the analysis of structural and functional neuromarkers of personality disorders whose main characteristic consists of persistent antisocial conduct, such as ASPD (De Brito et al., 2009) and psychopathy (Umbach et al., 2015), because they appear to be the most correlated to high rates of recidivism (Coppola, 2018). Research shows that these particular clinical populations share many traits, such as behavioral disinhibition or a lack of empathy, that are supposed to have common neurobiological bases (Coppola, 2018).

For example, abnormalities in limbic and paralimbic regions have been observed in individuals with psychopathic traits (Anderson and Kiehl, 2012) and impairments related to the prefrontal cortex are associated with disinhibition, emotional lability, and impulsivity (Chow, 2000; Yang and Raine, 2009).

Still, all such neurocriminological findings, obtained using conventional methods, do not enable us at this moment to make predictions of future risk. However, incorporating neurodata in A.I. prediction models appears to open up this possibility.

[1] A neural network is "a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes." [DARPA Neural Network Study (U.S.)., United States. Air Force. Systems Command., Lincoln Laboratory. (1989). DARPA neural *network study final report*. Lexington, Mass.: The Laboratory].

A first step toward A.I. prediction models using neuroimaging data is a study conducted by Aharoni et al. (2013), who used fMRI data to predict recidivism. The authors showed that activation in the dorsal anterior cingulate cortex (dACC), a brain region associated with impulse control and error processing, during a go/no-go task appeared to be associated with rearrest. The probability that offenders with relatively low anterior cingulate activity would be rearrested was approximately double compared to an offender with high activity in this region, keeping all the other risk factors constant. Low anterior cingulate activity, therefore, might be a potential neurocognitive biomarker for persistent criminal behavior (Aharoni et al., 2013).

Recently, a study by Kiehl et al. (2018) used machine learning coupled with neuroimaging to test whether brain age could help predict rearrest. Chronological young age is considered one of the key risk factors for recidivism. Young defendants are more likely to engage in risky behavior. Kiehl proposes that brain age is a better measure to account for individual differences than chronological age. The results of his study show that a predictive model involving neural measures of brain age performed better than previous models including only psychological and behavioral measures.

Even more recently, a study by Delfin et al. (2019) shows that improvements in recidivism prediction in forensic psychiatry might be possible by incorporating neuroimaging data into A.I. risk assessment models. The authors showed that the inclusion of resting-state regional cerebral blood flow (rCBF) measurements in an extended A.I. prediction model, containing neural measurements from eight brain regions, leads to an increase in predictive performance over traditional, empirical risk factors in a long-term follow-up of forensic psychiatric patients. Interestingly, they used 'classical' risk assessment *combined* with neuroimaging, which showed a better prediction in a forensic psychiatric population than the classical factors alone (Delfin et al., 2019).

In sum, preliminary findings in A.I. neuroprediction studies have produced some promising results. Still, the possible use of A.I. and 'brain-reading' in forensic populations raises several ethical and legal concerns, and the field of criminal justice should be cautious about their future use.

It is crucial to balance the preservation of offenders' individual rights on the one hand and the enhancement of public safety on the other.

PREDICTIVE ANALYSIS: TECHNOLOGICAL LIMITATIONS AND PITFALLS

Despite the opportunities previously discussed regarding the future possible use of A.I. neuroprediction techniques, several limitations should be considered; indeed, research about prediction tools and their successful application is still a challenging task (Poldrack et al., 2019).

This issue is well-known in the field of computational psychiatry, in which studies combining machine learning approaches and neuroimaging-based single subject prediction of brain disorders aim to classify patients with heterogeneous disorders (Arbabshirani et al., 2017; Bzdok and Meyer-Lindenberg, 2018). These studies, interestingly, reported varying degrees of accuracy (Neuhaus and Popescu, 2018), raising concerns about the methodology (Cearns et al., 2019). In fact, there is a need for best practices in predictive modeling (Poldrack et al., 2019); a problem of neuroprediction models is that, even though they can manage complex data such as brain imaging scans, they need best practices to ensure enough statistical power to test them (Varoquaux, 2018). Several issues deserve attention here.

First, application of neuroprediction techniques requires an inference from group-level to individual predictions (Hahn et al., 2017). Another challenge concerns validation of the results in a new group – different from the data set that was used to train the algorithm. The validity of prediction models is assessed by their ability to generalize; for most learning algorithms, the standard practice is to estimate the generalization performance through a process called 'cross-validation': the dataset is split into two sets, a training set, used to fit the model, and a test set (Hastie et al., 2009; Varoquaux, 2018), and subsets of the data are used to train and test the predictive performance of the model iteratively.

Notably, the use of cross-validation with small samples can lead to highly variable and inflated estimates of predictive accuracy (Luedtke et al., 2019; Poldrack et al., 2019). Training machine learning algorithms requires large amounts of data; using a limited sample size may cause so-called *overfitting*, in which the model fits perfectly to the specific data set used to train it, but fits poorly to new and unseen data (Hastie et al., 2009; Poldrack et al., 2019). There is still no agreement on the adequate size of the dataset (Cearns et al., 2019); Luedtke et al. (2019) recommend to perform prediction analyses with samples no smaller than several 100 observations. Acquiring many samples, however, is often difficult and costly, especially when neuroimaging data are involved (Arbabshirani et al., 2017).

ETHICAL AND LEGAL CHALLENGES

Prediction of recidivism using A.I. neuroprediction techniques evokes ethical and legal concerns, but also new possibilities. In what follows, we discuss some central ethical and legal issues.

First, we are confronted with the issue of *bias*. Since the advent of algorithmic risk assessment, a lot of reports have documented the fact that they are "dangerously" biased. The most famous case of supposed A.I. prejudice was reported by ProPublica in May 2016. COMPAS, an algorithm widely used in the US to guide sentencing by predicting the likelihood of a criminal reoffending, turned out to be racially biased against black defendants, according to ProPublica, because they were more likely than white defendants to be incorrectly classified as high risk ("false positives")[2] (Angwin et al., 2016). More recently, COMPAS has also been depicted as a "sexist algorithm" because its

[2]The company that produced the Compass algorithm, Northpointe, claimed in a report that the accuracy in the prediction of violence for both groups of defendants was the same: around 70% of crimes were predicted correctly (see Dieterich et al., 2016, COMPAS risk scales: demonstrating accuracy equity and predictive

algorithmic outcomes seem to systemically overclassify women in higher-risk groups (Hamilton, 2019). Similarly, Predpol, an algorithm designed to predict when and where crimes will take place, already in use in several US states, in 2016 – after an analysis of the Human Rights Data Analysis Group – was found to result in police *unfairly* targeting certain neighborhoods. Officers were repeatedly sent to areas of the city with a high proportion of people from racial minorities, regardless of the effective true crime rate in those areas (Ensign et al., 2018). Furthermore, facial recognition software, increasingly used in law enforcement, represents another potential source of both race and gender bias (Raji and Buolamwini, 2019*)*. Another example concerns Amazon's 'Rekognition' software, which is used by some police departments and other organizations. In 2018, the ACLU found that it incorrectly matched members of the Congress with people who had been charged with a crime, disproportionally misidentifying African-American and Latino members of Congress as the people in mug shots[3]. A recent study evaluating the accuracy of three commercial gender classifiers showed that they performed better in classifying male subjects than female subjects, and all of them performed worst on darker-skinned females (Buolamwini and Gebru, 2018). Moreover, recent studies show that, if left unchecked, word embeddings A.I. exhibit outdated gender stereotypes, such as "*doctors*" being male and "*receptionists*" being female (Bolukbasi et al., 2016).

These findings have led to a broader debate about the *fairness* of risk assessment using A.I. (Berk et al., 2018). Although algorithmic risk assessments can be perceived as a means of overcoming human bias, they could still reflect prejudice and institutionalized bias. A.I. is trained on data – for example, criminal files – that may themselves reflect biases on the part of police officers, prosecutors, or judges. Based on these data, the algorithm then "concludes" that groups with certain traits are more dangerous than others, while in fact, this is the result of biased data. This sometimes is referred to as "bias in-bias out." The results of A.I. prediction, in other words, highly depend on the quality of the data used. One advantage of using neuroimaging data – instead of police files – might be that neuroimaging does not reflect human bias. A.I. looks for correlations between brain activity and recidivism. Therefore, A.I. neuroprediction may offer possibilities to *decrease* bias in risk assessment. However, also since neuroprediction may be incorporated in existing risk assessment tools (see the study by Delfin et al., 2019), bias will remain a problem as long as there is no solution to bias in algorithms in general.

Furthermore, we should keep in mind that risk assessment is "quintessentially discriminatory" (Binns, 2017), meaning that it is all about classifying subjects into groups of low or high-risk individuals based on group traits. Neuromarkers for recidivism will undoubtedly be more prevalent in certain groups than in others. Treating groups of people differently because of their "brain" raises difficult questions about what constitutes unjustified unequal treatment. This question, however, is not typical of A.I. neuroprediction, but is a central issue in risk assessment and fairness in general (Nadelhoffer et al., 2012, p. 95; Tonry, 2014). Classifying people into groups based on their brain scan, even if useful to prevent possible harms, could easily lead to stigmatization and discriminating effects for those considered "high risk" in other aspects of the individual's life. It could become a sort of *modern phrenology*, by discriminating between people based on what their brain looks like. While certain institutional procedures could discriminate against those considered "high risk," stigmatization could be a more social process that excludes certain individuals based on their risk profile; for instance, stigmatization may be a consequence of sex offenders' registration (Tewksbury, 2005).

A second point concerns *privacy*. The neurodata and other data used to predict recidivism can clearly also be of interest for other purposes. For instance, for insurance companies, or when screening job applicants. Who should have access to these data, and under which conditions? Should insurance companies have access to them, and if not, should they be able to request such a procedure in order to assess the risk of a particular candidate client? Clearly, in this case, data protection – and possible access – is a fundamental issue, already highly debated in algorithms used in the era of big data. Obviously, there is also a parallel with the current debate on the nature of consent and the degree of control citizens have regarding health information in biobanks. The discussion of commercialization of genetic/health information and rights of control ("biorights") are likely to intensify in the coming years (see also Caulfield and Murdoch, 2017).

A third, related point concerns the probability of a *negative 'self-fulfilling prophecy*.' This qualm comes from recent studies, showing that receiving genetic risk information can actually *influence* your behavior, physiology, and subjective experience and change your overall risk profile (Turnwald et al., 2019). Researchers from Stanford University found that when people were told of a genetic tendency for either obesity or lower exercise capacity, acquiring this information had a physiological impact on their bodies, modifying how they responded to a meal or to exercise. A persistent discovery was that perceptions of risk altered health outcomes, therefore those informed of having the high-risk gene had a worse outcome than those informed of having the protective one (Turnwald et al., 2019). Following these findings, one may wonder how the mindset of people may be affected when you inform them about their own risk information, either genetic or neural, and how this could actually alter their risk profile. This shows that providing information may also require ethical and/or legal research and regulation.

Furthermore, it is still not clear how to exactly classify and conceptualize neurodata as risk factors. For example, in a study by Kiehl et al. (2018), a measure of brain age (gray matter) is used to predict recidivism. Chronological age is often considered a static factor, but when referring to brain measures,

parity. Retrieved from www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html). The different levels of false positives among black defendants and white defendants were to be attributed, according to Northpointe, to different base rates in the prevalence of crime among black and white defendants. It is possible to have the algorithm acquire the same level of false positives over groups with a different base rate. However, this comes at the cost of reduced accuracy. There is an extensive literature on fairness in A.I. prediction, and its trade-offs (Berk et al., 2018). The text about these algorithms is partially based on Cossins (2018).

[3]https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

we should reflect on how they should be conceptualized among risk factors. For instance, given the plasticity of the brain, should we consider brain age as a dynamic or static risk variable? How do we evaluate an offender if, for example, brain age and normal age differs, and how would this modify his/her neuroprediction profile? If we consider neurodata as dynamic factors, and, as such, available to be modified through interventions, we could talk, instead of in terms of a pure "prediction," in terms of targets for treatment and other intervention types. Used in this way, neuroprediction could help to prevent crime through more individualized correctional and socio-rehabilitative measures, and could also enable offenders to return to the community sooner. As in "personalized medicine" – a therapeutic approach in which an individual's genetic and epigenetic information is used to tailor drug therapy or preventive care[4] – neuroprediction could help to target interventions to the individual's "needs."

There is another effect of the emphasis on prediction that is relevant here. Currently, A.I. is used in the criminal justice system, mainly to predict recidivism. A.I. risk assessment typically does not offer a causal model of crime and therefore, is not designed to show opportunities to intervene and to mitigate risk (Berk, 2019, pp. 17–18). Barabas et al. (2018) conclude: "when risk assessments are used primarily as a predictive technology, they fuel harmful trends toward mass incarceration and growing inequality in the justice system."

We should acknowledge that A.I. neuroprediction in the first place merely establishes correlations between brain images and the risk of recidivism. However, if it is indeed possible to develop interventions based on neurodata, this might offer offenders an opportunity to avoid incarceration (Nadelhoffer et al., 2012, pp. 85–86). This could be possible because, different from historical data and other risk variables, like a person's demographic characteristics such as ethnicity, age, and gender, that cannot be changed, neurodata hold the potential to become targets for new rehabilitative interventions and prevention programs, aiming to reduce exposure to risk factors for psychopathic traits and preventing at-risk individuals from engaging in criminal behavior later in life (Ling and Raine, 2018).

This is particularly important since the prison environment may have negative effects on neurocognitive functioning. In fact, studies found that incarceration might lead to reduced self-control (Meijers et al., 2018). Still, the possibility of intervention also entails its own ethical and legal issues: for an offender, it may be hard to choose between a deprivation of liberty and undergoing (possibly somewhat invasive) treatment, especially in light of the right to refuse medical treatment (Meynen, 2018). However, this again is not a problem that is typical of interventions based on "A.I. neuroprediction."

A fourth, and related, issue concerns *consent and coercion*; if and when these techniques will be fully developed and are ready to be used, there may be a possibility of performing cognitive liberty violations forcing people to undergo scans without consent for sentencing or punitive purposes (Ligthart, 2019; Meynen, 2019). Coercion, both technical and ethical or legal, not only relates to the force used, because not all the imaging techniques allow for this, but also to their use within the context of a threat or an offer that cannot be refused (Meynen, 2017). One way to counter this issue is to strictly regulate informed consent for neuroprediction tests.

Fifth, we should take into account something called the "seductive allure" that neuroimaging exerts on courts. Juries and judges apparently tend to overestimate the accuracy of neuroscientific evidence, and, although neuroimaging aims to reduce uncertainty and to increase the objectivity in forensic settings, the use of neuroimaging in courts is at risk of being misleading, due to cognitive biases in the evaluation of evidence (Scarpazza et al., 2018). Introducing neuroprediction could therefore lead to some overreliance on neurodata.

Furthermore, machine learning algorithms are considered to be '*black-boxes of decision-making*'; the way in which they perform decisions is not fully comprehensible to stakeholders, and not even to expert data scientists (London, 2019; Pedreschi et al., 2019). In addition, we have to be cautious about what is called the "the control problem"; i.e., the tendency of human operators to become complacent with machines, devolving responsibility and becoming over-reliant on the outputs of autonomous systems, even when they are biased (Pedreschi et al., 2019). In order to avoid overreliance, it seems important for A.I. systems to be transparent: it should be possible to explain to judges and a jury how they produce their results (Gunning and Aha, 2019), and stakeholders should be capable to appropriately trust and manage these tools, reasoning on how a specific output is given and on the basis of what rationale (Pedreschi et al., 2019). Even if this is actually complicated by the fact that most risk assessment algorithms are proprietary, it seems important for society that A.I. algorithms can be made intelligible, in order to be accountable for their decisions (Weld and Bansal, 2019).

Of note, legal systems may have criteria for the admissibility of scientific evidence in the courtroom. For instance, in the US legal context *Daubert* and *Frye* are used as standards. As we do not focus on specific legal systems, we will not go into this in more detail, but clearly such legal criteria would be relevant for courtroom use of new technologies (Shats et al., 2016).

Moreover, it is important to make a decision about the required accuracy of these technologies. Current risk assessment tools often have an AUC of about 0.70 (Douglas et al., 2017); is that enough for such algorithms, or should the threshold be higher, like 0.80 or 0.90? These are normative choices that have to be made before deciding to allow the use of this kind of technology to prevent crime.

Additionally, we need to consider the lack, at present, of a 'true' prediction model; a limitation of the papers previously discussed is that, instead of talking about 'pure' prediction, they can be classified as *postdiction* studies; postdiction generally relates to retrospectively making an assertion or deduction about an event based on information available after the event (Yamada et al., 2015) but, as applied to the context of statistical models, the distinction between prediction and postdiction is about whether the assessment of the model's success involves the same data as were used to build the model or new data not used

[4]https://www.nature.com/subjects/personalized-medicine

in model construction (Gauch and Zobel, 1988; Hastie et al., 2009). Research suggests that models for predictive applications, such as biomarkers, require larger sample sizes than standard statistical approaches (Varoquaux, 2018). Furthermore, in the studies discussed before, data about neuromarkers of recidivism have been collected after the commission of crimes, so we cannot establish when brain differences observed developed (Cope et al., 2014). A future challenge is to develop a true prediction model, able to identify those at the highest risk for committing crimes, and research in neuroimaging coupled with A.I. may be the key in developing such model.

Finally, there appears to be a more remote problem, looming on the horizon. Suppose that these A.I. algorithms – either with or without brain imaging – become really good predictors, wouldn't that introduce a form of determinism we have not witnessed before? The A.I. system may be considered to have some "divine" foreknowledge about what will happen, which may have negative effects on the freedom people experience and exert. A belief in free will seems to have positive effects (Crescioni et al., 2016; Feldman et al., 2016).

Still, the more pressing concern nowadays is that we are not quite good at predicting risk – even with A.I. – and that we nonetheless often apply sanctions based on the supposed dangerousness of the offender. If A.I. becomes more accurate with the help of neuroimaging, it could reduce the number of persons incorrectly classified as high risk and can therefore reduce sanctions that in fact are not legitimate, helping to interrupt the so-called "cycles of crime" (Barabas et al., 2018).

CONCLUSION

There is still a way to go before combined neuroscience and AI-based violence risk assessment tools can be implemented in the criminal justice system. Still, A.I. is already being used in criminal justice systems. Because of the far-reaching consequences of these type of technologies – and also given some rapid developments in recent years – it is important to consider ethical and legal concerns. Besides discussing technological limitations and pitfalls of predictive analysis, we identified six key issues deserving attention: dealing with bias, privacy, the possibility of a 'self-fulfilling prophecy,' coercion and consent, the allure of neuroimaging data and the need for A.I. systems to be explainable. Finally, we pointed to the more remote issue of how highly accurate predictions might introduce a form of determinism we have not witnessed before – but this is still far away.

Still, we would like to emphasize that accurate risk prediction is extremely valuable for both safety and justice reasons. Therefore, in principle, we argue that technologies that may be helpful in this respect should at least be explored, and if ready, used in criminal justice and forensic psychiatry. In addition, neuroprediction and A.I. bring their own, in a way new, ethical and legal challenges, and we will have to deal with them – preferably before the technologies are used. More specifically, we have to find solutions to prevent systems from reflecting our own human biases in order to enable them to provide objective and trustworthy data.

Therefore, we argue that the use of AI-based systems in criminal justice and forensic psychiatry should be subjected to substantial regulation to protect citizens from system errors or misuse. On such basis, we highlight the importance of accurate harms/benefits analyses not only when these technologies will be fully available, but also while they are being researched and developed.

AUTHOR CONTRIBUTIONS

LT, GM, and SF conceived the content of the manuscript and wrote and revised the manuscript. LT drafted the manuscript. JB and ET wrote and revised the manuscript. All authors read and approved the final manuscript.

REFERENCES

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* 8:14. doi: 10.3389/fninf.2014.00014

Aharoni, E., Vincent, G. M., Harenski, C. L., Calhoun, V. D., Sinnott-Armstrong, W., Gazzaniga, M. S., et al. (2013). Neuroprediction of future rearrest. Proceedings of the national academy of sciences of the united states of america. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6223–6228. doi: 10.1073/pnas.1219302110

Anderson, N. E., and Kiehl, K. A. (2012). The psychopath magnetized: insights from brain imaging. *Trends Cogn. Sci.* 16, 52–60. doi: 10.1016/j.tics.2011.11.008

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). *Machine Bias*. New York, NY: ProPublica.

Arbabshirani, M. R., Plis, S., Sui, J., and Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *Neuroimage* 145(Pt B), 137–165. doi: 10.1016/j.neuroimage.2016.02.079

Barabas, C., Virza, M., Dinakar, K., Ito, J., and Zittrain, J. (2018). "Interventions over predictions: reframing the ethical debate for actuarial risk assessment," in *Proceedings of FAT conference (FAT 2018). ACM*, New York, NY, 62–76.

Berk, R. (2019). *Machine Learning Risk Assessments in Criminal Justice Settings*. Berlin: Springer.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2018). Fairness in criminal justice risk assessments: the state of the art. *arXiv.org* [Preprint], doi: 10.1177/0049124118782533

Berk, R., and Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. *Fed. Sentenc. Rep.* 27, 222–228. doi: 10.1525/fsr.2015.27.4.222

Binns, R. (2017). Fairness in machine learning: lessons from political. *Philos. Proc. Mach. Learn. Res.* 81, 1–11.

Blitz, M. J. (2017). "Lie detection, mind reading, and brain reading. in: searching minds by scanning brains," in *Palgrave Studies in Law, Neuroscience, and Human Behavior*, (Cham: Palgrave Macmillan), 45–58. doi: 10.1007/978-3-319-50004-1_3

Bode, S., and Haynes, J.-D. (2009). Decoding sequential stages of task preparation in the human brain. *Neuroimage* 45, 606–613. doi: 10.1016/j.neuroimage.2008.11.031

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv. Neural Inform. Proc. Syst.* 20, 4349–4357.

Bray, S., Chang, C., and Hoeft, F. (2009). Applications of multivariate pattern classification analyses in developmental neuroimaging of healthy and clinical populations. *Front. Hum. Neurosci.* 3:32. doi: 10.3389/neuro.09.032.2009

Buolamwini, J., and Gebru, T. (2018). "Gender shades: intersectional accuracy disparities in commercial gender classification," in *Proceeding of F.A.T*, New York, NY.

Bzdok, D., and Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatry* 3, 223–230.

Caulfield, T., and Murdoch, B. (2017). Genes, cells, and biobanks: yes, there's still a consent problem. *PLoS Biol.* 15:e2002654. doi: 10.1371/journal.pbio.2002654

Cearns, M., Hahn, T., and Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Transl. Psychiatry* 9:271. doi: 10.1038/s41398-019-0607-2

Chow, T. W. (2000). Personality in frontal lobe disorders. *Curr. Psychiatry Rep.* 2, 446–451. doi: 10.1007/s11920-0000031-5

Conroy, M. A., and Murrie, D. C. (2007). *Forensic Assessment Of Violence Risk: A Guide For Risk Assessment And Risk Management*. Hoboken, NJ: John Wiley & Sons Inc, doi: 10.1002/9781118269671

Cope, L. M., Ermer, E., Gaudet, L. M., Steele, V. R., Eckhardt, A. L., Arbabshirani, M. R., et al. (2014). Abnormal brain structure in youth who commit homicide. *Neuro. Clin.* 4, 800–807. doi: 10.1016/j.nicl.2014.05.002

Coppola, F. (2018). Mapping the brain to predict antisocial behaviour: new frontiers in neurocriminology, 'new'challenges for criminal justice. *U.C.L. J. Jurisprud. Spec.* 1, 106–110.

Cossins, D. (2018). Discriminating algorithms: 5 times AI showed prejudice. *New Scientist* (accessed January 10, 2019).

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI)brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261–270. doi: 10.1016/s1053-8119(03)00049-1

Crescioni, A. W., Baumeister, R. F., Ainsworth, S. E., Ent, M., and Lambert, N. M. (2016). Subjective correlates and consequences of belief in free will. *Philos. Psychol.* 29, 41–63.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D. G., Acharyya, M., Loughead, J. W., et al. (2005). Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage* 28, 663–668. doi: 10.1016/j.neuroimage.2005.08.009

De Brito, S. A., Mechelli, A., Wilke, M., Laurens, K. R., Bartoli, A. J., Barker, G. J., et al. (2009). Size matters: increased grey matter in boys with conduct problems and callous-unemotional traits. *Brain* 132(Pt 4), 843–852. doi: 10.1093/brain/awp011

Delfin, C., Krona, H., Andine', P., Ryding, E., Wallinius, M., and Hofvander, B. (2019). Prediction of recidivism in a long-term follow-up of forensic psychiatric patients: incremental effects of neuroimaging data. *PLoS One* 14:e0217127. doi: 10.1371/journal.pone.0217127

Dieterich, W., Mendoza, C., and Brennan, T. (2016). *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*. Northpointe Inc.

Douglas, K. S., and Kropp, P. K. (2002). A prevention-based paradigm for violence risk assessment: clinical and research applications. *Crim. Just. Behav.* 29, 617–658. doi: 10.1177/009385402236735

Douglas, T., Pugh, J., Singh, I., Savulescu, J., and Fazel, S. (2017). Risk assessment tools in criminal justice and forensic psychiatry: the need for better data. *Eur. Psychiatry* 42, 134–137. doi: 10.1016/j.eurpsy.2016.12.009

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018). "Runaway feedback loops in predictive policing," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Berlin.

Fazel, S., Singh, J. P., Doll, H., and Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis. *B.M.J.* 345:e4692. doi: 10.1136/bmj.e4692

Feldman, G., Chandrashekar, S. P., and Wong, K. F. E. (2016). The freedom to excel: belief in free will predicts better academic performance. *Pers. Indiv. Differ.* 90, 377–383.

Fink, B. C., Steele, V. R., Maurer, M. J., Fede, S. J., Calhoun, V. D., and Kiehl, K. A. (2016). Brain potentials predict substance abuse treatment completion in a prison sample. *Brain Behav.* 6:501. doi: 10.1002/brb3.501

Gauch, H. G., and Zobel, R. W. (1988). Predictive and postdictive success of statistical analyses of yield trials. *Theoret. Appl. Genetics* 76, 1–10. doi: 10.1007/BF00288824

Gendreau, P., Little, T., and Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: what works. *Criminology* 34, 575–608. doi: 10.1111/j.1745-9125.1996.tb01220.x

Gunning, D., and Aha, D. (2019). DARPA's explainable artificial intelligence (X.A.I.) Program. *A.I. Magaz.* 40, 44–58. doi: 10.1609/aimag.v40i2.2850

Hahn, T., Nierenberg, A., and Whitfield-Gabrieli, S. (2017). Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol. Psychiatry* 22, 37–43. doi: 10.1038/mp.2016.201

Hamilton, M. (2019). The sexist algorithm. *Behav. Sci. Law* 37, 145–157. doi: 10.1002/bsl.2406

Hart, S. D. (1998). The role of psychopathy in assessing risk for violence: conceptual and methodological issues. *Legal Criminol. Psychol.* 3, 121–137. doi: 10.1111/j.2044-8333.1998.tb00354.x

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Berlin: Springer, doi: 10.1007/978-0-387-84858-7

Haynes, J.-D. (2012). "Brain reading," in *I Know What You're Thinking: Brain imaging and Mental Privacy*, eds D. Sarah, G. Rees, and J. L. Sarah, (Oxford: Oxford University Press), doi: 10.1093/acprof:oso/9780199596492.003.0003

Haynes, J. D., and Rees, G. (2006). Neuroimaging: decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534.

Haynes, J. D., Sakai, K., Rees, G., Gilbert, S., Frith, C., and Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Curr. Biol.* 17, 323–328.

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *Neuro. Clin.* 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Horikawa, T., Tamaki, M., Miyawaki, Y., and Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science* 340, 639–642. doi: 10.1126/science.1234330

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* 8, 679–685. doi: 10.1038/nn1444

Kehl, D., Guo, P., and Kessler, S. (2017). *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Responsive Communities*. Available online at: http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041 (accessed March 1, 2019).

Kiehl, K. A., Anderson, N. E., Aharoni, E., Maurer, J. M., Harenski, K. A., Rao, V., et al. (2018). Age of gray matters: neuroprediction of recidivism. *Neuroimage* 19, 813–823. doi: 10.1016/j.nicl.2018.05.036

Kim, J., Calhoun, V. D., Shim, E., and Lee, J. H. (2016). Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *Neuroimage* 124(Pt A)), 127–146. doi: 10.1016/j.neuroimage.2015.05.018

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain* 131(Pt 3), 681–689. doi: 10.1093/brain/awm319

Koutsouleris, N., Borgwardt, S., Meisenzahl, E. M., Bottlender, R., Möller, H. J., and Riecher-Rössler, A. (2012). Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: results from the FePsy study. *Schizophr. Bull.* 38, 1234–1246. doi: 10.1093/schbul/sbr145

Lemm, S., Benjamin, B., Thorsten, D., and Mueller, K. (2011). Introduction to machine learning for brain imaging. *Neuroimage* 56, 387–399. doi: 10.1016/j.neuroimage.2010.11.004

Ligthart, S. (2019). "Coercive neuroimaging technologies in criminal law in Europe: exploring the implications for the prohibition of ill-treatment (article 3 ECHR)," in *Regulating New Technologies In Uncertain Times Information Technology And Law Series*, ed. L. Reins, (Berlin: Springer), 83–102. doi: 10.1007/978-94-6265-279-8-6

Ling, S., and Raine, A. (2018). The neuroscience of psychopathy and forensic implications. *Psychol. Crim. Law* 24, 296–312. doi: 10.1080/1068316X.2017.1419243

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hast. Center Rep.* 49, 15–21. doi: 10.1002/hast.973

Luedtke, A., Sadikova, E., and Kessler, R. C. (2019). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clin. Psychol. Sci.* 7, 445–461. doi: 10.1177/2167702618815466

Meijers, J., Harte, J. M., Meynen, G., Cuijpers, P., and Scherder, E. J. A. (2018). Reduced self-control after 3 months of imprisonment. A pilot study. *Front. Psychol.* 9:69. doi: 10.3389/fpsyg.2018.00069

Meynen, G. (2017). Brain-based mind reading in forensic psychiatry: exploring possibilities and perils. *J. Law Biosci.* 4, 311–329. doi: 10.1093/jlb/lsx006

Meynen, G. (2018). Forensic psychiatry and neurolaw: description, developments and debates. *Int. J. Law Psychiatry.* 65:101345. doi: 10.1016/j.ijlp.2018.04.005

Meynen, G. (2019). Ethical issues to consider before introducing neurotechnological thought apprehension in psychiatry. *AJOB Neurosci.* 10, 5–14. doi: 10.1080/21507740.2019.1595772

Monahan, J., and Skeem, J. L. (2015). Risk Assessment in Criminal Sentencing (September 17, 2015). Annual Review of Clinical Psychology, Forthcoming; Virginia Public Law and Legal Theory Research Paper, No. 53. Available online at SSRN: https://ssrn.com/abstract=2662082 (accessed January 10, 2019).

Morse, S. J. (2015). *Neuroprediction: New Technology, Old Problems. Faculty Scholarship at Penn Law.1619*. Available online at: https://scholarship.law.upenn.edu/faculty_scholarship/1619 (accessed January 10, 2019).

Nadelhoffer, T., Bibas, S., Grafton, S., Kiehl, K. A., Mansfield, A., Sinnott-Armstrong, W., et al. (2012). Neuroprediction, violence, and the law: setting the stage. *Neuroethics* 5, 67–99. doi: 10.1007/s12152-010-9095-z

Neuhaus, A. H., and Popescu, F. C. (2018). Sample size, model robustness, and classification accuracy in diagnostic multivariate neuroimaging analyses. *Biol. Psychiatry* 84:e0081-2.

Ombao, H., Lindquist, M., Thompson, W., and Aston, J. (2017). *Handbook of Neuroimaging Data Analysis*. New York: Chapman and Hall/CRC.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., and Turini, F. (2019). Meaningful explanations of black box ai decision systems. *Proc. AAAI Conf. Artif. Intellig.* 33, 9780–9784.

Poldrack, R. A., Huckins, G., and Varoquaux, G. (2019). Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry.* 27:2019. doi: 10.1001/jamapsychiatry.2019.3671

Poldrack, R. A., Monahan, J., Imrey, P. B., Reyna, V., Raichle, M. E., Faigman, D., et al. (2018). Predicting violent behavior: what can neuroscience add? *Trends Cogn. Sci.* 22, 111–123. doi: 10.1016/j.tics.2017.11.003

Raji, I., and Buolamwini, J. (2019). "Actionable auditing: investigating the impact of publicly naming biased performance results of commercial a.i. products," in *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society*, New York, NL.

Rubbert, C., Mathys, C., Jockwitz, C., Hartmann, C. J., Eickhoff, S. B., Hoffstaedter, F., et al. (2019). Machine-learning identifies Parkinson's disease patients based on resting-state between-network functional connectivity. *Br. J. Radiol.* 2019:20180886. doi: 10.1259/bjr.20180886

Scarpazza, C., Ferracuti, S., Miolla, A., and Sartori, G. (2018). The charm of structural neuroimaging in insanity evaluations: guidelines to avoid misinterpretation of the findings. *Transl Psychiatry.* 8:227. doi: 10.1038/s41398-018-0274-8

Sen, B., Borle, N. C., Greiner, R., and Brown, M. (2018). A general prediction model for the detection of ADHD and Autism using structural and functional M.R.*I. PloS one* 13:e0194856. doi: 10.1371/journal.pone.0194856

Shats, K., Brindley, T., and Giordano, J. (2016). Don't ask a neuroscientist about phases of the moon: applying appropriate evidence law to the use of neuroscience in the courtroom. *Cambridge Quarterly of Healthcare Ethics* 25, 712–725. doi: 10.1017/S0963180116000438

Shen, G., Horikawa, T., Majima, K., and Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS. Comput. Biol.* 15:e1006633. doi: 10.1371/journal.pcbi.1006633

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., and Dean, K. (2014). International perspectives on the practical application of violence risk assessment: a global survey of 44 countries. *Int. J. Foren. Ment. Health.* 13, 193–206. doi: 10.1080/14999013.2014.922141

Steele, V. R., Fink, B. C., Maurer, J. M., Arbabshirani, M. R., Wilber, C. H., Jaffe, A. J., et al. (2014). Brain potentials measured during a Go/NoGo task predict completion of substance abuse treatment. *Biol. Psychiatry* 76, 75–83. doi: 10.1016/j.biopsych.2013.09.030

Steele, V. R., Maurer, J. M., Arbabshirani, M. R., Claus, E. D., Fink, B. C., Rao, V., et al. (2018). Machine learning of functional magnetic resonance imaging network connectivity predicts substance abuse treatment completion. *Biol. Psychiatry* 3, 141–149. doi: 10.1016/j.bpsc.2017.07.003

Steele, V. R., Rao, V., Calhoun, V. D., and Kiehl, K. A. (2015). Machine learning of structural magnetic resonance imaging predicts psychopathic traits in adolescent offenders. *Neuroimage* 145(Pt B), 265–273. doi: 10.1016/j.neuroimage.2015.12.013

Tewksbury, R. (2005). Collateral consequences of sex offender registration. *J. Contemp. Crim. Just.* 21, 67–81. doi: 10.1177/1043986204271704

Tonry, M. (2014). Legal and ethical issues in the prediction of recidivism. *Fed. Senten. Rep.* 26, 167–176.

Turnwald, B. P., Goyer, J. P., Boles, D. Z., Silder, A., Delp, S. L., and Crum, A. J. (2019). Learning one's genetic risk changes physiology independent of actual genetic risk. *Nat. Hum. Behav.* 3, 48–56.

Umbach, R., Berryessa, C., and Raine, A. (2015). Brain imaging research on psychopathy: implications for punishment, prediction, and treatment in youth and adults. *J. Crim. Just.* 43, 295–306.

Varoquaux, G. (2018). Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* 180(Pt A), 68. doi: 10.1016/j.neuroimage.2017.06.061

Weld, S. D., and Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 70–79. doi: 10.1145/3282486

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. (2017). Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* 28, 4136–4160. doi: 10.1093/cercor/bhx268

Yamada, Y., Kawabe, T., and Miyazaki, M. (2015). Awareness shaping or shaped by prediction and postdiction: editorial. *Front. Psychol.* 6:166. doi: 10.3389/fpsyg.2015.00166

Yang, Y., and Raine, A. (2009). Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: a meta-analysis. *Psychiatry Res.* 174, 81–88. doi: 10.1016/j.pscychresns.2009.03.012

Yip, S. W., Scheinost, D., Potenza, M. N., and Carroll, K. M. (2019). Connectome-based prediction of cocaine abstinence. *Am. J. Psychiatry* 176, 156–164. doi: 10.1176/appi.ajp.2018.17101147

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 24 January 2020
doi: 10.3389/fpsyg.2019.02926

Assessing Risk Among Correctional Community Probation Populations: Predicting Reoffense With Mobile Neurocognitive Assessment Software

Gabe Haarsma[1†], Sasha Davenport[1†], Devonte C. White[1,2], Pablo A. Ormachea[1], Erin Sheena[1] and David M. Eagleman[1,3]*

[1] The Center for Science and Law, Houston, TX, United States, [2] Administration of Justice Department, Texas Southern University, Houston, TX, United States, [3] Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, United States

OPEN ACCESS

Edited by:
Eric García-López, Instituto Nacional de Ciencias Penales, Mexico

Reviewed by:
Vincenzo Tigano, University of Magna Graecia, Italy
Andrea Lavazza, Centro Universitario Internazionale, Italy

***Correspondence:**
David M. Eagleman
davideagleman@stanford.edu

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *27 September 2019*
Accepted: *11 December 2019*
Published: *24 January 2020*

Citation:
Haarsma G, Davenport S, White DC, Ormachea PA, Sheena E and Eagleman DM (2020) Assessing Risk Among Correctional Community Probation Populations: Predicting Reoffense With Mobile Neurocognitive Assessment Software. Front. Psychol. 10:2926. doi: 10.3389/fpsyg.2019.02926

We seek to address current limitations of forensic risk assessments by introducing the first mobile, self-scoring, risk assessment software that relies on neurocognitive testing to predict reoffense. This assessment, run entirely on a tablet, measures decision-making via a suite of neurocognitive tests in less than 30 minutes. The software measures several cognitive and decision-making traits of the user, including impulsivity, empathy, aggression, and several other traits linked to reoffending. Our analysis measured whether this assessment successfully predicted recidivism by testing probationers in a large urban city (Houston, TX, United States) from 2017 to 2019. To determine predictive validity, we used machine learning to yield cross-validated receiver–operator characteristics. Results gave a recidivism prediction value of 0.70, making it comparable to commonly used risk assessments. This novel approach diverges from traditional self-reporting, interview-based, and criminal-records-based approaches, and can also add a protective layer against bias, while strengthening model accuracy in predicting reoffense. In addition, subjectivity is eliminated and time-consuming administrative efforts are reduced. With continued data collection, this approach opens the possibility of identifying different levels of recidivism risk, by crime type, for any age, or gender, and seeks to steer individuals appropriately toward rehabilitative programs. Suggestions for future research directions are provided.

Keywords: risk assessment, machine learning, neurolaw, predictive validity, neurocognitive

INTRODUCTION

The Bureau of Justice Statistics estimates that 12–13 million people are processed annually through jail facilities nationwide – and that 68% of released felony-level prisoners are rearrested within 3 years, 79% within 6 years, and 83% within 9 years (Alper et al., 2018). The criminal justice system has long seen the value in determining the best course of treatment, sentencing, or release of an offender by administering tests or reviewing records to roughly classifying individuals in terms

of future risk of rearrest (Berk, 2017). However, concerns about the fairness and accuracy of risk assessments (Eckhouse et al., 2019) have increased the stakes that prosecutors and judges face when making risk-based determinations. The development of actuarial risk assessments sought to address inadequacies and provide statistical soundness to the approach of using only clinical judgment and criminal history (Sreenivasan et al., 2000). It follows that a statistically sophisticated tool which could uncover predictive traits and dynamic factors at the individual level, while filtering out unfair biases, used in conjunction with clinical judgment, would be beneficial to persons in the system and society.

Forecasting an individual's likelihood of future criminality has been part of the criminal justice system "since judges have been judging" (Gottfredson, 1987; Berk and Hyatt, 2015). Methodologies have expanded the scope of assessing risk of reoffending, and over the past 40 years courts have become significantly more advanced in attempting to divide high- from low-risk offenders. The predicted level of risk can be used to determine pretrial release, steer bail amount (Desmarais and Lowder, 2019), length of sentence, or probation status, and it can also shed light on rehabilitation strategies. Having an idea of the risk someone poses to the public can allow courts to more optimally produce sentences to balance freedoms against societal protection.

Individuals are quite different in their predispositions (Eagleman, 2011), and because lives are complex, and crime is contextual, there will never be a test that accurately predicts the future (such as the "pre-cogs" in the movie *Minority Report*); nonetheless, risk assessments have been shown to perform at a higher rate of accuracy than subjectivity of psychiatrists and parole boards (Grove et al., 2000; Ægisdóttir et al., 2006; Spohn, 2008; Dressel and Farid, 2018). There are over 60 risk assessments used to specifically measure recidivism in the United States (Barry-Jester et al., 2015; Casselman and Goldstein, 2015). In this paper, we present 17 of them (**Table 1**) to establish a landscape of the most widely used tests.

Most risk assessments ask questions that measure factors that are classified as static or dynamic (Austin, 2004; Desmarias, 2013). Early risk assessments that used prior arrest records or interview-based assessments predominantly focused on static factors – that is, variables that cannot be changed (race, place of birth, or arrest record). Some such static factors are used regularly to inform rehabilitation tracks, such as gender, age, and crime type. On the other hand, some researchers are concerned that static factors yield a risk score that is too restrictive, because such factors do not allow for the possibility that an individual can change.

By contrast, traits that can change over time (called dynamic factors) offer more indication of an offender's current and future behavior (Andrews and Dowden, 2007; Ward and Fortune, 2016). These include factors such as education, employment, marital status, and cognitive traits. Dynamic risk factors can be mitigated with intervention strategies (Bonta and Andrews, 2007). Some researchers have developed assessment tools that combine static and dynamic factors to estimate the likelihood of reoffending and offer appropriate recommendations. This provides correctional professionals with a baseline for determining risk while allowing for change over time as well.

TABLE 1 | Commonly used risk assessments, their stated purpose, and their median area under the curve (AUC) (Singh et al., 2011; Desmarais et al., 2016).

Risk assessment	AUC	Purpose
COMPAS	0.67	General and violent recidivism, pretrial misconduct
IRAS	0.63	General recidivism
LSI-R	0.64	General recidivism
ORAS	0.66	General recidivism
PCRA	0.71	Post-conviction reoffense, under supervision
PSA	0.66	Pretrial risk assessment
RMS	0.67	General recidivism
SARA	0.70	Domestic violence
SAVRY	0.71	Violent risk in youth
SORAG	0.75	Sex offender
SPIn-W	0.73	Gender-responsive (for women)
Static-99	0.70	Sex offenders, pre-release
STRONG	0.74	General recidivism
SVR-20	0.78	Sexual violence
TRAS	0.67	General recidivism
VRAG	0.74	Violent risk
WRN	0.67	General recidivism

COMPAS, Correctional Offender Management Profile for Alternative Sanction; IRAS, Indiana Risk Assessment Survey; LSI-R:SV, Level of Service Inventory; ORAS, Ohio Risk Assessment Survey; PCRA, Federal Post Conviction Risk Assessment; PSA, Public Safety Assessment; RMS, Risk Management System; SARA, Spousal Assault Risk Assessment; SAVRY, Structured Assessment of Violence Risk in Youth; SORAG, Sex Offender Risk Appraisal Guide; SPIn-W, Service Planning Instrument–Women; Static-99; STRONG, Static Risk and Offender Needs Guide; SVR-20, Sexual Violence RIsk-20; TRAS, Texas Risk Assessment Survey; VRAG, Violence Risk Appraisal Guide; WRN, Wisconsin Risk and Needs.

The most commonly used actuarial risk assessments ask similar questions about the individual, and share similar predictive strength as measured by the receiver operating characteristic curve (ROC curve), and the area under the curve (AUC), which ranges from 0.5 (no predictability) to 1 (perfect predictability). This value serves as an evaluation metric of how good the models are at distinguishing between two classes. Models are built to make probability predictions about each participant's chance of falling into two classes: those who will recidivate, and those who will not. The higher the ROC AUC is, the better it is at classifying between the two groups.

The best assessments range in AUC values from the mid-0.6s to mid-0.7s. Some risk assessments measure risk for specific crimes, or populations, such as STATIC-99 and SAVRY (Structured Assessment of Violence Risk in Youth), which concentrate specifically on sex crimes and risk of violence for youth, respectively. However, note that some studies indicate that having a high predictive value for some measures (such as sex crimes) correlates with low predictive value for other serious crimes (Langton et al., 2007). Thus, these more specific risk assessments are narrower in their predictive capacities.

While current risk assessments have been successful, they also have limitations. First, the information used to score

the assessments is generally gathered from a single criminal offense level and may not be flexible enough to apply to another. For example, if a risk assessment is validated to score well at the felony level, it is not guaranteed to be accurate at the misdemeanor level (Pope-Sussman and Turner, 2015). Second, in the absence of expensive, ongoing training, the variance between rater scores can be a concern (Lowenkamp et al., 2004; Duwe, 2017). Third, actuarial risk assessments can be time consuming, affecting key stakeholders such as administration, practitioners, and test takers (Desmarias, 2013). Fourth, there may be problems in taking an assessment that was validated at one point of the criminal justice pipeline and using it in a different application for which it may not be as fair and accurate, e.g., pre-trial vs. recidivism (Dressel and Farid, 2018). Further, recent studies and lawsuits have highlighted the possibility of racial bias when an assessment relies on subjectivity of the interviewer, as well as using static and dynamic factors that correlate with race (e.g., education and previous criminal history) (Harcourt, 2015; Dressel and Farid, 2018). Although race is not included in risk assessments, many factors included in risk assessments correlate heavily with race. In a 2014 speech to the National Association of Criminal Defense Lawyers, former Attorney General Eric Holder warned that sentencing decisions based on "immutable characteristics may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society" (Holder, 2014). Recent court rulings have further highlighted the unfairness of the use of proprietary scoring algorithms that do not allow one to see how the score was calculated, and thus assess its accuracy or contest the score (State v. Loomis, 2016; Kehl et al., 2017).

To address these limitations, we have developed an innovative assessment tool for predicting reoffense using rapid, interactive tests based on standard neuropsychological tests (Ormachea et al., 2017; **Figure 1**). The NeuroCognitive Risk Assessment (NCRA) measures key criminogenic factors (attentiveness, aggression, risk seeking, empathy, future planning, emotional processing, and impulsivity), all of which have been identified in the literature as cognitive traits linked to reoffending. We then used machine learning models to quantify an individual's risk for re-offense, which yields findings significantly better than using general linear modeling alone.

There are several benefits to the NCRA. The test is self-administered on a mobile device (such as an iPad), and test administrators require no training to supervise individuals taking the test. The interactive battery is "gamified", making the test interactive and engaging compared to traditional questionnaires (**Table 1**). Administrators are not required to have extensive training, or a professional degree to interpret the results, and they do not need to directly administer the assessment to participants individually while they are taking the test, thus allowing it to be taken in a group setting. The test is self-administered by the participant, and the visual and audible instructions, along with the practice rounds, make the battery easy to understand, even among low literacy populations (the current version requires only a 4th grade reading level). Further, the minimal text in the games is easily translated into different languages, allowing testing in different tongues and locations. Collectively, these qualities make adoption of the NCRA more accessible, scalable, affordable, and less time/resource consuming than traditional assessments.

Analysis of NCRA scores is based on machine learning and therefore can be grouped with the most current risk assessment tools as an actuarial method (Berk and Hyatt, 2015). It can inform case management by allowing the ongoing tracking of decision-making traits while a person participates in programs, which can be useful in case planning or identifying levels of service, needs, or risk management. For example, knowing an offender's aggression and impulsivity profile could assist case workers in monitoring progress through treatment programs and target more effective behavioral therapies. Rather than using static factors such as criminal history or demographics – which often come under scrutiny for potential bias – the NCRA measures dynamic cognitive factors

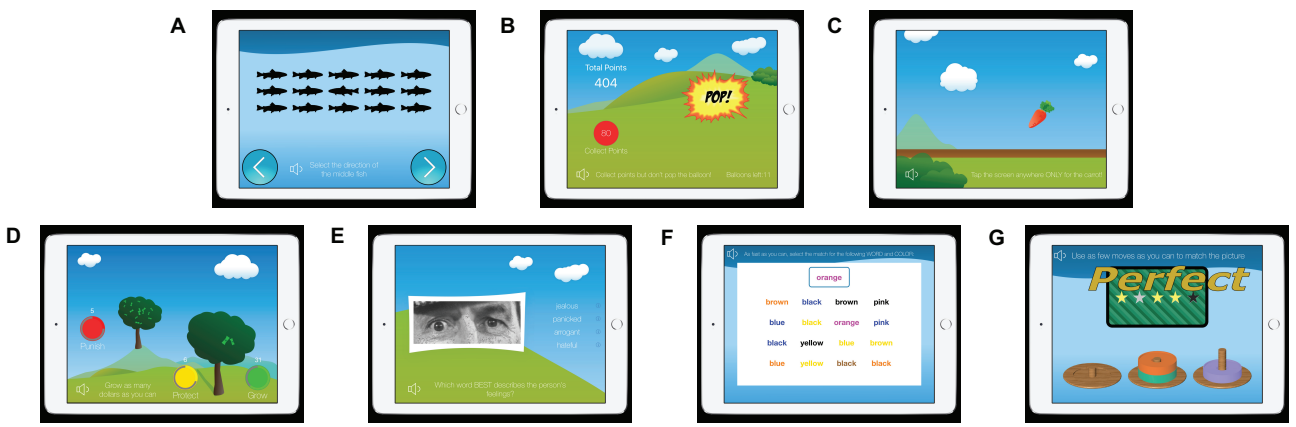


FIGURE 1 | Screenshots of the NeuroCognitive Risk Assessment (NCRA): **(A)** the Eriksen Flanker, **(B)** Balloon Analog Risk Task, **(C)** Go/No-Go, **(D)** Point-Subtraction Aggression Paradigm, **(E)** Reading the Mind Through the Eyes, **(F)** Emotional Stroop, and **(G)** Tower of London (Ormachea et al., 2017).

in decision-making such as risk taking, aggression, empathy, impulsivity, and attention – all of which are dynamic traits that can be improved.

Predictive validity is a spectrum, in which scores estimate the likelihood of a person recidivating. Assessed over a population, an assessment can be measured by its AUC value, where 0.5 depicts no discriminative ability, 1.0 represents perfect prediction, and most good risk assessments have an AUC of around 0.70 (Howard, 2016). The most commonly used risk assessments have virtually the same predictive validities, so there is little in terms of a hierarchy of effectiveness between these different risk assessments (Campbell et al., 2009; Yang et al., 2010; **Table 1**). This is because many of the most commonly used risk assessments use essentially the same factors, self-reporting and weighted questions about the individual past and current position, and only differ in the way they analyze those factors (Monahan and Skeem, 2014), typically with proprietary algorithms that cannot be studied. Because they are using the similar risk factors and have the similar AUC predictive validity scores, the risk assessments might have reached a "glass ceiling" that cannot be broken (Monahan and Skeem, 2014). In other words, to attain a more accurate risk assessment, the input predictive factors must differ from using *only* interviews, self-reporting questionnaires, or records-based assessments, as well as having a defined outcome (Fazel et al., 2012). Utilizing cognitive traits of the NCRA offers inputs that allow new predictive features to emerge, ones that can also inform rehabilitation needs. It bears re-emphasizing that it is impossible for risk assessment to predict whether someone will recidivate with 100% accuracy, because people are unpredictable, life is complex, and crime can be contextual. Nonetheless, in the same way that life insurance companies improve their returns by building actuarial tables to assess whether a customer is high-, medium-, or low-risk, perfect predictability is not necessary to improve sentencing and rehabilitation decisions in the criminal justice system.

In this study, we use machine learning to examine the predictive validity of the NCRA in a forensic community corrections population. The literature in computer science has demonstrated that machine learning statistics can forecast more accurately than previous approaches based on regression analysis (Trevor et al., 2009; Berk and Hyatt, 2015; Duwe and Rocque, 2017). These machine learning techniques differ from the typical statistical analysis used in conventional risk assessments that often have anticipated and weighted relationships between crime and recidivism that are then built into the model. By contrast, machine learning models finds relationships within the data that are not prescripted, and may not be obvious, like interactions or non-linear relationships, but nonetheless increase predictive validity and give a near-optimal prediction of recidivism (Berk and Hyatt, 2015).

Moreover, the NCRA eliminates factors that discriminate against individuals based on race and socioeconomic status, because the assessment does not need to compute data that are linked to these characteristics. Rather, the NCRA uses only neurocognitive measures (which assess attributes linked with criminality and reoffense) that can be modified and improved.

MATERIALS AND METHODS

Study Design

Informed consent was obtained from 730 probationers who volunteered to participate in the study. Participants self-administered the tablet-based test in about 30 min. All participants were provided headphones for the auditory and visual test instructions. They reviewed instructions and played brief unscored practice tests prior to each assessment to ensure they understood each test's rules. Participants were not offered any reward or compensation for participating in the study, and were debriefed afterward.

Once consent was obtained, the mobile device was handed to the participant. The battery began with a short customizable questionnaire for participants to enter demographics or answer questions relevant to the program or facility they were in. Each test began with video-based, audible instructions, using language and text designed for a 4th grade reading level. Non-scored practice rounds were offered after the instructions, which included feedback to ensure the test-taker fully understood the instructions. The NCRA is comprised of seven tests (**Figure 1**), each of which was selected based on their relationship to reoffense, as detailed in the neurocognitive literature (Ormachea et al., 2017). The tests were then "gamified" (to benefit engagement and attention) and optimized to balance data collection against rapid testing time.

The seven tests, as deployed in conjunction with one another, does not exist anywhere else, so the NCRA is unique, despite growing from tests that have been historically used to analyze neurocognitive behavior. Each test lasts 2–4 min and (depending on the speed of the test taker) the entire battery is administered in about 30 min. After the participant completes each test, scores are automatically calculated by the software. The results are stored with HIPAA-compliant security in the cloud.

To analyze results, NCRA test data were taken for each participant, along with age, gender, and current offense category (**Tables 2**, **3**) participants were charged with at the time of testing. Two publicly available criminal history databases were used to ensure the most accurate information. Information from the state obtained through the county probation department was used to track reoffenses, or any arrest that happened post assessment.

Participants

The NCRA was self-administered by 730 participants in the Harris County Community Supervision and Corrections Department. We tracked reoffense of participants by utilizing two data sources: the Harris County District Clerk public criminal records database and the Texas Department of Public Safety. The earliest check on rearrest was conducted at 4 months post-assessment, and re-checked regularly up to 2 years post-assessment. Recidivism is defined by any subsequent arrest after the initial arrest (Andrews and Bonta, 1995; Markman et al., 2016). Technical violations of conditions of probation (e.g., failing to update current address or missing an appointment) was not counted, even if the event resulted in adjusted probation terms.

TABLE 2 | Overview of probationers (age and gender) and recidivism, by current offense category.

Category	*N*	Recidivate (*N*)	Recidivate (%)	Gender (Male%)	Age (Median)	Age (SD)
DWI	182	15	8.2	78.0	31.9	10.7
Drug	187	38	20.3	78.6	28.4	10.0
Non-violent	35	13	37.1	85.7	29.1	9.3
Property	122	23	18.9	59.8	27.7	10.2
Sexual non-violent	8	1	12.5	75.0	29.4	8.9
Sexual violent	8	1	12.5	87.5	38.6	10.1
Violent	188	35	18.6	77.1	27.9	8.6
Total	730	126	17.3	75.3	28.8	10.0

TABLE 3 | Self-reported race/ethnicity and number of previous arrests.

		Race/ethnicity (%)				Arrests (*N*)						
Category	*N*	Asian	Black	Hispanic	White	Other	0	1	2	3	4–10	11+
DWI	182	3.3	23.1	41.8	27.5	4.4	10	73	41	27	31	0
Drug	187	3.2	32.1	34.8	22.5	7.5	5	49	40	50	41	2
Non-violent	35	0.0	65.7	28.6	2.9	2.9	2	8	9	9	7	0
Property	122	1.6	44.3	31.1	22.1	0.8	8	46	26	20	22	0
Sexual non-violent	8	0.0	50.0	25.0	25.0	0.0	0	3	2	1	1	1
Sexual violent	8	0.0	25.0	25.0	25.0	25.0	0	5	1	1	1	0
Violent	188	1.1	44.7	32.4	16.5	5.3	11	53	42	48	31	3
Total	730	2.2	36.8	34.8	21.2	4.9	36	237	161	156	134	6

Note that none of this self-reported information is used in any of the machine learning sections.

The testing group comprised adult participants recruited from the Harris County Community Services and Corrections Department (CSCD) from 2017 to 2019. Of the participants, 550 (75.3%) were male and 180 (24.7%) were female. Participants had either been assigned to probation through the court from a previous arrest, or were in pre-trial assessments for a recent arrest. Participants were charged with misdemeanors (332) or felonies (398). Descriptive statistics of participants population are provided in **Tables 2**–**4**.

In 2 years, 126 of the 730 participating probationers (17.3%) recidivated in Harris County (**Table 2**). This is an underestimation of the actual recidivism of the offenders, as some crime goes undetected (for example, as happens in jurisdictions we do not have access to). Also, Class C misdemeanors (as defined by the Texas State penal code) were left out of this analysis – e.g., crimes that result in no jail time and have fines <$500.

About the Assessment

Throughout the development of the NCRA, our aim has been to determine how underlying cognitive traits (and specifically, those that have established links to criminal behavior) can be used to harvest insights into recidivism. An appreciation of how these decision-making traits are linked to reoffending can optimize individualized sentencing strategies, and can steer rehabilitative program recommendations toward individualized treatment (Ormachea et al., 2016). We've leveraged neuropsychological tests that are sensitive to different cognitive domains, gamified them, and time-optimized them. By running them on a tablet, accuracy and reaction time (down to the millisecond scale) can inform scoring (Zelazo et al., 2014). The following is a brief description of the tests:

The *Eriksen Flanker task* is a focus and attention task that measures executive functioning. A school of fish that is heading left or right is displayed on the screen. The middle fish may point in the same direction (*congruent*) or a different direction than the school (*incongruent*). The object of the game is to press an arrow on the screen indicating the direction that only the middle fish is facing, ignoring all other distracting fish.

The *Balloon Analog Risk task* (*BART*) was chosen to measure risk taking behavior. The object is to inflate the balloon by pressing on it, as much as you dare, earning points as the balloon continues to grow. But beware, the balloon could burst at any time and all points "risked" for that trial will be lost.

The *Go/No-Go* (*GNG*) *task* measures a participant's ability to inhibit impulsivity. The aim is to touch the screen as fast as possible when a carrot is plucked up from the ground. However, in a fraction of trials, an eggplant pops up instead of a carrot, and in this circumstance the user is meant to inhibit the urge to tap.

The *Point-Subtraction Aggression Paradigm* (*PSAP*) is a test that measures reactive aggression. The aim is to grow dollars on the tree as fast as possible by rapidly tapping the "grow" button. However, a second player (who is actually the computer) is trying to grow money on their tree as well, and will sometimes "steal" dollars from the participant's tree, resulting in two more choices that appear: the participant can protect the dollars they've grown so far, or retaliate and "punish" player 2 by eliminating one of their dollars. Either of those choices requires multiple taps and takes time away from the participants clear instructions, which is

TABLE 4 | Self-reported education and employment.

Education	*N*	(%)	Employment	*N*	(%)
Middle or junior high	17	2.3	Not employed	152	20.8
Some high school	143	19.6	Homemaker	14	1.9
High school or GED	323	44.2	Student	47	6.4
Some/in college	154	21.1	Employed	380	52.1
College graduate	36	4.9	Student/employed	12	1.6
Graduate school	7	1.0	Other	125	17.1
Vocational school	50	6.8			

Note that education and employment information is not used in any of the machine learning sections.

to "grow" as much money as they can. The game measures how aggressively a participant is prone to react to a slight.

The *Reading the Mind Through the Eyes* (*RTMTE*) *task* measures social cognition, specifically, empathy, which tends to be deficient in violent offenders (Baron-Cohen et al., 1997; Domes et al., 2013; Seidel et al., 2013). Users are presented with the upper half of a face. They are tasked with selecting a word (out of four words) that best describes the face's emotional state. Of 30 trials, we track the number that are incorrect.

The *Emotional-Stroop test* detects microsecond apprehensions related to the negatively charged words. Consisting of several levels, the user starts with a traditional Stroop test involving words and colors (neutral, congruent, incongruent). The final levels introduce series of neutral words, positive words, and negatively charged words (related to drug use), in the same format. The tests pick up on delays of negative words that require more emotional processing time and indicate a relationship.

The *Tower of London* (*TOL*) is a shape- and color-matching game that tests the ability to simulate future consequences and plan ahead. The user is shown three pegs that up to three colored discs are stacked on. The task is to make the fewest moves possible to match the pattern of disks shown.

After the participants completed the NCRA, we regularly queried databases for evidence of recidivism, as well as the crime category and level (misdemeanor vs. felony). For this, we used the Harris County District Clerk public criminal records database and the Texas Department of Public Safety criminal history search. The former was automatically queried monthly, and provides detailed records about offenses from Harris County. This was augmented with additional arrest and court data from the Texas Department of Public Safety, which was queried semi-annually for offenses committed outside Harris County.

Features Used and Feature Sets

Each neurocognitive test generates raw unstructured data, called features. These typically involve the individual trial number, specific information about the trial, millisecond resolution timing when a button was pressed, and whether the individual answer or action was correct or incorrect. From the raw unstructured data, machine learning features were developed for each test. These are generally a set of summary statistics for the test (**Table 5**). Beyond these features, participants' age and gender demographics, as well as the current offense category were used. We never included other variables, such as race/ethnicity, education, or employment.

By analyzing the NCRA feature data alone and then combining the NCRA with basic information about the participant (age, gender, and current offense category), we filtered for the most predictive NRCA features and introduced four different feature sets (**Table 6**).

Machine Learning and Predictive Validity

Traditionally, the most suitable method for estimating and predicting the probability of an event at a single point in time is a standard linear logistic regression (generalized linear models or glm). To take advantage of newly developed advances in data science, we applied machine learning methods to optimize the feature selection and address possible non-linearities in the data. Machine learning has a number of benefits over traditional statistical methods, including efficient handling of noisy data, non-linearities, and numerous predictors, as well as being able to automatically mine and estimate complex interactions (Tollenaar and van der Heijden, 2013). To build on this, we report the findings of both the generalized linear regression along with machine learning packages (**Table 7**).

Predictive validity describes the degree to which a score predicts a criterion measure – in our case, recidivism. To assess the predictive validity of the machine learning methods chosen, the focus of this paper is on the ROC curve, which plots the true positive rate (sensitivity) against the false positive rate (1-specificity) for every possible cut-off threshold. An ROC curve captures the predictive ability of a binary classifier system (in this case, *recidivates* or *does not recidivate*). When using such a plot, one traditionally measures the AUC, which gives the probability that any given classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (Hanley and Mcneil, 1982). A perfect model which completely separates the two classes would have 100% sensitivity and specificity, which will result in an AUC of 1. In contrast, a completely ineffective model would result in a ROC curve that closely follows the diagonal line and would have an area under the ROC curve of approximately 0.5 (Kuhn and Johnson, 2013). We measure the NCRA via the ROC AUC, as this is the most widely used method in measuring predictive validity in risk assessments (Rice and Harris, 2005; Singh et al., 2013). The higher the AUC, the more accurate the model is in predicting (Cortes and Mohri, 2003; Clémençon and Vayatis, 2010).

TABLE 5 | Definitions of the machine learning features used in each NCRA test.

Feature	Description
Eriksen Flanker	
Time median	Median response time
Time standard deviation	Standard deviation response time
Exec effect	Median congruent trials – median incongruent trials
Frac correct	Percent of correct trials
NIH score	National Institute of Health Flanker score
Balloon analog risk task	
Pop	Number of popped balloons
Time collected (*)	Total time/points collected from unpopped balloons
Pressed time median	Median time/points collected
Pressed count median	Median number of balloon inflate presses
Duration time median	Median time/duration of inflate presses
Go/no-go	
Correct go	Correct number of Go's (carrot)
Correct no go	Correct number of No-Go's (eggplant)
Time correct go	Mean response time of correct Go's
Point-subtraction aggression paradigm	
Grow (*)	Number of individual grow taps/50
Protect ratio	Protect taps/all taps
Punish ratio (*)	Punish taps/all taps
Reading the mind through the eyes	
Correct (*)	Number of correct trials
Time median (*)	Median response time
Time standard deviation	Standard deviation response time
Dict lookup	Number of trials any trial word is looked up
Emotional Stroop	
Test correct	Test round with feedback number of correct trials
Test time (*)	Test round with feedback mean response time
Black correct	Std Stroop color words in black number of correct trials
Black time (*)	Std Stroop color words in black mean response time
Con color correct	Std Stroop color words congruent color number of correct trials
Con color time	Std Stroop color words congruent color mean response time
Incon color correct	Std Stroop color words incongruent color number of correct trials
Incon color time (*)	Std Stroop color words incongruent color mean response time
Neutral correct	Neutral words number of correct trials
Neutral time	Neutral words mean response time
Pos Neg correct	Positive and negative words number of correct trials
Pos Neg time (*)	Positive and negative words mean response time
Tower of London	
Solved	Number of trials solved
Aborted (*)	Number of trials aborted (giving up)
All moves	Number of total moves
Dup moves (*)	Number of duplicated moves
Extra moves	Number of extra moves to solve
Illegal moves (*)	Number of illegal moves
Mean time	Mean trial time
Solved mean time	Mean trial time for solved trials
Solved median time	Median trial time for solved trials
First move time	Mean time waited before moving a disk in a trial
First move frac (*)	Mean fraction of time waited before moving a disk in a trial
Final time	Time the last disk was moved
Test moves	Number moves in the test round
Test time	Time spend in the test round
Test solved	Was the test round solved
Disk speed	Mean time between start and stop of moving a disk

Asterisks signify the most predictive features.

TABLE 6 | Feature sets defined, as used in machine learning modeling analysis.

Feature set	Description
Full NCRA	NCRA test data, without any other information
RFE NCRA	Recursive feature elimination producing the top 13 most predictive features of the NCRA, with no other information
Full NCRA + Demographics	NCRA test data combined with demographics (age, gender, and the current crime category at the time of testing)
RFE NCRA + Demographics	Recursive feature elimination producing the top 13 most predictive features of the NCRA combined with demographics

The Challenges of Small and Unbalanced Data Sets in Machine Learning

Large datasets are better for machine learning, so additional care has to be taken when working with smaller sets. Despite our dataset being sizable when compared to traditional risk assessment validation studies (Singh et al., 2011), the data are used for both the development/training of the model and the validation of it. The first concern is the overfitting of data, which can lead to low errors in training, but high variance when using the model on a hold-out validation test. This error gets amplified when using high dimensional, noisy data (such as human behavior) that drowns out the nuances and leads to poorly generalizable results.

Many of the machine learning algorithms and models that are used require one or more tuning parameters to be set that have a large effect on their performance. In each type of model we must select the model that performs best in a range of tuning parameters. To select the best model over its range within the estimation set, we use five times repeated 10-fold cross-validation. The repeated cross validation (RCV) will select the optimal set of tuning parameters for a given machine learning algorithm. After optimal model parameters are established, 80% of the data is used to construct a final model. The other 20% is used as a hold-out set which is used to evaluate the performance of the final candidate model.

Both the model training and hold-out validation set are randomly split while balancing the recidivism class. A single training and validation run has significant variability in predictive validity from the random split. To address this, the splitting training and validation are repeated up to 100 times with different random seeds for the split. This will create a distribution from which to get an average, the results of the repeated samples represent the validity of the model's predictions, which leads to a ROC, which is then used to create an AUC.

A dataset is said to be unbalanced when the class of interest (minority class) is much rarer than normal behavior (majority class). The overall recidivism rate is 17.3% (our minority class), which creates a mild imbalance in the data. This kind of mild imbalance is not a problem for most machine learning algorithms. However, when combined with the small overall data set, it is possible that some machine learning algorithms can become sensitive. To overcome the class imbalance problem it is possible to oversample the minority class or use a more advanced sampling method like SMOTE (synthetic minority over-sampling technique; Chawla et al., 2002).

Feature selection is primarily focused on removing non-informative or redundant predictors from the model. Many machine learning methods will estimate parameters for every term in the model. Because of this, the presence of non-informative variables can add uncertainty to the predictions and reduce the overall performance of the model. As a first pass, a filter is used to remove features that are highly correlated. Secondly, recursive feature elimination (RFE) is employed to find the set of most informative features (Kuhn and Johnson, 2013).

Machine Learning Algorithms and Software Used

Classification models were built utilizing the CARET package (short for Classification And REgression Training) version 6.0.84 package in the R programming language (version 3.6.1; R Core Team, 2019).

A long list of available machine learning algorithms is supported by the CARET package. For this analysis, the predictive performances of commonly used machine learning techniques with a binary outcome variable were used (**Table 7**). Overviews of the various machine learning methods can be found in Kuhn and Johnson (2013) and Tollenaar and van der Heijden (2013).

RESULTS

We will successively present the predictive validity of the NCRA for general recidivism models with ROC curves and AUCs. We first look at four different feature sets and seven machine learning methods (**Table 7**) in order to make an overall judgment on the performance of the NCRA. The ROC curves that follow are used as a quantitative assessment of the machine learning methods and the feature sets for which the AUC are summarized into a single AUC number.

Table 8 shows the AUCs for each combination of previously described feature sets and machine learning algorithms. Overall, the Glmnet and LDA algorithms performed similarly well, with the former producing slightly higher AUC across every feature set. ROC curves in **Figure 2** show all machine learning methods for the RFE NCRA + Demographics feature set, which corresponds with the reported AUCs in the last row of **Table 8**.

The GLM with ridge and lasso regularization (Glmnet) machine learning method performs the best overall for each feature set. On average LDA, GBM, SVM, and GLM performed second highest with very little difference between the methods. The lowest performing machine learning method is k-NN, which is unsurprising given the simplicity of the method. In general, observations on the machine learning methods are in line with the findings of Tollenaar and van der Heijden (2013, 2019).

TABLE 7 | Machine learning models used with corresponding R statistical analysis package.

Label	Method	R package
GLM	Generalized linear models	stats version 3.6.1
LDA	Linear discriminant analysis	MASS version 7.3-51.4
k-NN	k-Nearest neighbors	class version 7.3-15
SVM	Support vector machines (polynomial)	kernlab version 0.9-27
GMB	Generalized boosted modeling	gbm version 2.1.5
RF	Random forest	ranger version 0.11.2
Glmnet	GLM with ridge and lasso regularization	glmnet version 2.0-18

We next explored whether AUCs improved when we addressed class imbalance by oversampling the minority class. For all machine learning methods (except SVM) using the SMOTE method to correct the imbalance did not improve the predictive validity. At a minority imbalance of 17.3% the imbalance is insufficient to cause problems for most machine learning methods.

Receiver operating characteristic curves in **Figure 3** show the performance of the various feature sets for the Glmnet machine learning method, which also corresponds with the reported AUCs in the last column of **Table 8**. Whether demographics were included or not, feature sets that eliminate features without predictive power (RFE sets) perform better: on average, the AUCs increased by 0.02. The RFE technique is a useful algorithm to identify which individual features are stronger predictors, exclude the poorer performing features, and focus the machine learning algorithms on those reduced sets.

We next found that adding in general information about the participant age, gender, and the current crime category (**Table 2**) slightly enhanced the predictive performance of the models. This is not surprising, given the established literature on gender in crime, the age–crime curve, and differing recidivism rates for different crime categories. On average, the model produced an AUC 0.04 higher when we included this information; however, the amount of improvement depended on the machine learning method: there was little improvement for SVM, and the most for GLM, LDA, and Glmnet.

Receiver operating characteristic curves in **Figure 4** show the combination of machine learning algorithm *Glmnet* with the *RFE NCRA + Demographics* feature set. The blue line is the average of 100 runs with different splits between training and validation sets, and each individual run is shown by a black line. The amount of variability between individual runs is due to the relatively small sample size for machine learning, and will reduce as the

TABLE 8 | ROC curve AUCs by feature sets along with machine learning algorithms used.

Feature sets	GLM	LDA	k-NN	SVM	GBM	RF	Glmnet
Full NCRA	0.60	0.61	0.56	0.64	0.63	0.60	0.64
RFE NCRA	0.64	0.65	0.58	0.66	0.64	0.62	**0.66**
Full NCRA + Demographics	0.65	0.66	0.59	0.65	0.66	0.63	0.69
RFE NCRA + Demographics	0.68	0.69	0.60	0.67	0.67	0.66	**0.70**

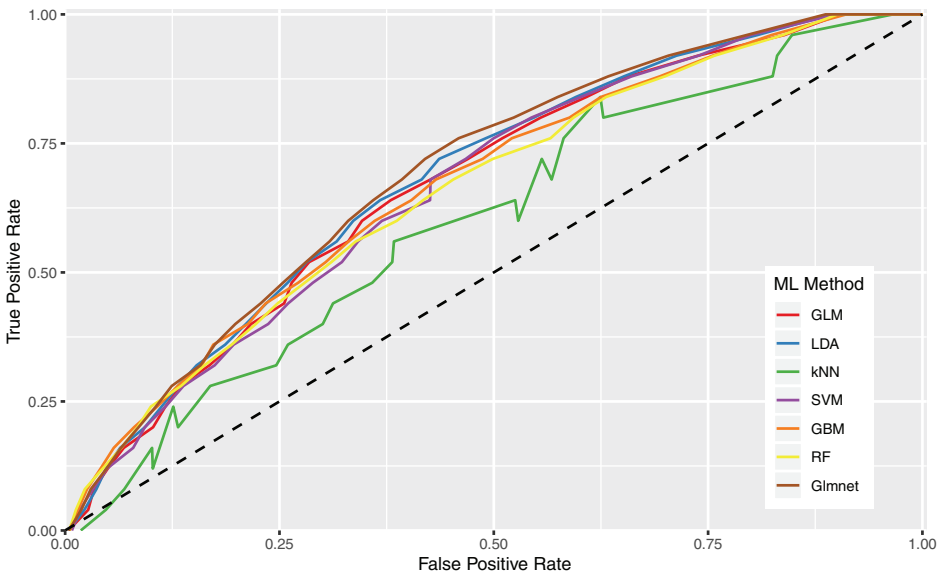


FIGURE 2 | Receiver operating characteristic curves illustrating predictive performance of all machine learning algorithms when looking at the RFE NCRA + Demographics feature set.

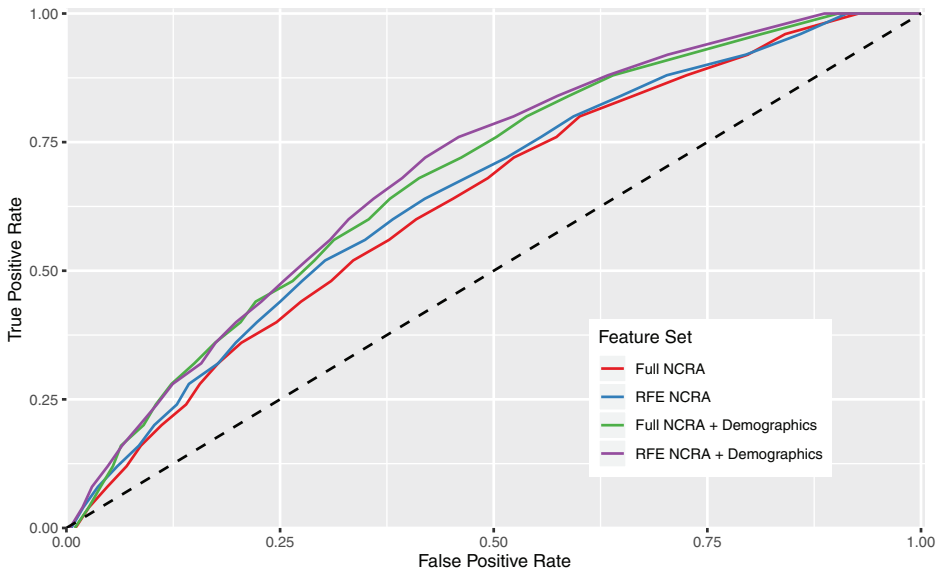


FIGURE 3 | Receiver operating characteristic curves illustrating predictive performance of the Glmnet machine learning method over all feature sets.

sample size increases. The average ROC produces an AUC of 0.70 which falls between the uppermost bin of the "good" category and "excellent" category for predictive performance indicators (Desmarais and Singh, 2013).

In line with our hypothesis, the feature set of *Full NCRA* with no other information and using a *Glmnet* machine learning algorithm has a relatively competitive AUC that only dipped slightly to 0.66, which is comparable to reported predictive validity in commonly used risk assessments (**Table 1**), and ranked in the "good" category (Desmarais and Singh, 2013). Because this analysis is based on a small sample size for the purposes of presenting preliminary results, further study is needed to (1) confirm that the AUC increases with more data points (Berk and Bleich, 2013) and (2) allow more time for participants' possible re-arrests.

Finally, we hope to be able to subset by crime type in the near future, as our data pool expands. As an exploratory step, we split our data into felony and misdemeanor crime level. Despite

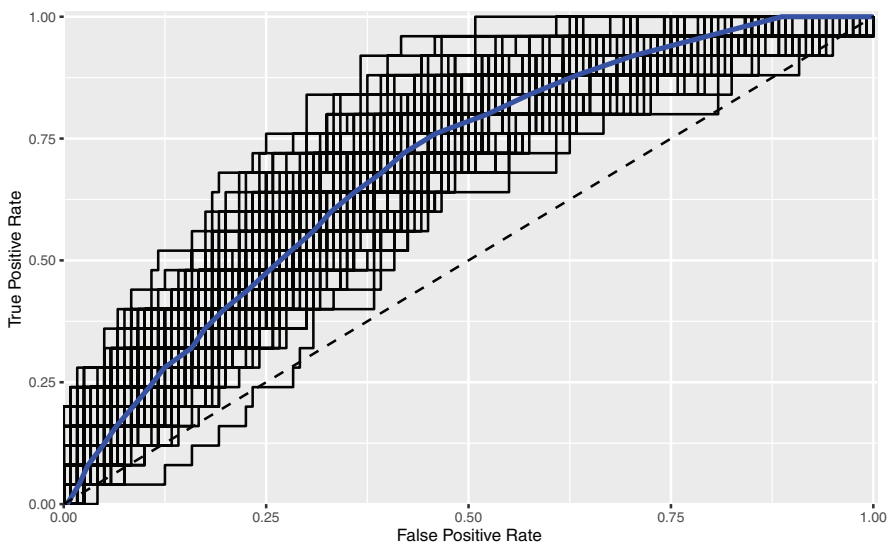


FIGURE 4 | Receiver operating characteristic curves illustrating predictive performance of the Glmnet machine learning method for the RFE NCRA + Demographics feature set.

a reduction in sample size from splitting the data, there appears to be promise that the NCRA will be able to produce predictive scores for general recidivism in either category. Preliminary results show an AUC for felony level crimes at 0.72, a moderate to strong effect with an AUC for misdemeanor level crimes showed an AUC of 0.68. Next steps will include further exploration of felony versus misdemeanor, comparing violent with non-violent crime levels, as well as other detailed level sub-crime types.

Strengths and Limitations

This study is the first to show that neurocognitive tests optimized as games on a mobile tablet has a predictive accuracy, measured by AUC, comparable to commonly used risk assessments. Further, by using neurocognitive testing – with no racial information – we are able to address important critiques about implicit and explicit racial bias. Further, our approach also avoids the possible confounds of static factors (e.g., using distant criminal history to inform future behavior). Additionally, as the development of the NCRA proceeds, we will adopt an algorithmic equity checklist (Osoba et al., 2019) to minimize any undesirable equity outcomes.

The NCRA is a flexible tool in terms of administration, implementation, and utility. By combining neurocognitive tests with existing actuarial assessment protocols, the benefit of a deeper understanding of deviant decision-making can be factored into sentencing and treatment programs. There is also room for expanding the test into other areas of cognitive testing by adding new tests and conducting additional feature exploitation to carve out the most predictive variables.

We note several data limitations. Our sample size is on the low side for machine learning models; nonetheless, it is high enough to maintain a stable predictive model over multiple runs. As new data are added, we expect the predictive validity will increase. However, no matter how good our test gets in the future, note that no predictive test will ever approach perfection: life and behavior are simply too complex for that.

Another limitation is the crime level distribution of prison-eligible offenders. Felony-level violent offenders and recidivists are represented in the sample; however, participants in this analysis were offenders who committed crimes that were eligible for probation or pre-trial diversion. Despite using two databases to track rearrests both locally and in border counties, the study is limited by siloed jurisdictional databases, which undercount arrest rates for all participants who may have gone on to commit crime in other states.

Also, it is possible that data from this probation sample in Houston may not generalize well to other jurisdictions, and results may differ at key points of the criminal justice pipeline (e.g., pretrial versus pre-sentencing).

DISCUSSION

The purpose of this study was to analyze the predictive accuracy as measured by AUC, of recidivism in a community probation population using gamified neurocognitive testing. Our results demonstrate that a rapid, gamified, test on a tablet computer can perform as well (or, in many cases better) than the most commonly used risk assessments.

As the use of risk assessments has grown, so has the scrutiny of their efficacy, methods, and purpose, especially in light of equality in machine learning algorithms (Eaglin, 2017; Berk et al., 2018). Following current trends, it is likely that an increasing number of states will mandate risk assessments for defendants and offenders. Recent legislation has been drafted to adopt such policies on a large scale, such as the Pre-trial Integrity and Safety Act. This proposed legislation drew from the implementation of risk assessments in Kentucky to expand such programs in the United States (Harris and Paul, 2017). The current work adds an instrument to the toolbox that can deliver both large-scale social and direct economic impacts.

The NCRA offers a variety of benefits with a predictive validity comparable to widely used risk assessments. This is the first tool to assess the underlying neurocognitive drivers of decision-making in a criminal justice setting. The NCRA has the potential to become a time- and resource-saving option for arraignment assessments. Improvements in predicting re-offense have the potential to translate into a safer society by more effectively modulating sentencing and steering rehabilitative strategies.

Our next steps will be aimed at studying and deploying the NCRA at key points of the criminal justice continuum. It would be beneficial to test individuals at other points of the system, including pre-trial to help determine bail options, probation assessment, jail, and prison intake when rehabilitation programs are assigned. At the end of the pipeline, we're interested in exploring testing re-entry programs, parole supervision, and explore juvenile justice pipelines. With greater variability in testing timepoints, and also with population, a richer picture can emerge of the trajectory of decision-making.

Outside of the courts, the assessment may help assist in determining beneficial diversion, reentry, or community-based programs for individuals reentering society post conviction, which would call for applying customized thresholds. With a cognitive/behavioral snapshot of an individual, we hope to be better able to address individual need for each person to receive the rehabilitation programs they need to succeed.

To ensure that we do not introduce address racial bias, we did not use any information about race, number of previous arrests, education, or employment in our machine learning models. The best-performing model did include general information such as age, gender, and crime level information. However, we are able to show the small gap between using those factors (0.70 AUC) and using cognitive performance alone (0.66 AUC) showing promise that with a larger sample size, we can drop age, gender, and crime level all together when appropriate (Skeem et al., 2016; Douglas et al., 2017) – yielding a test that only uses neurocognitive measures to predict reoffense with higher AUCs than standard risk assessments.

Previous psychometric evaluation of the NCRA suggested that predictors may exist that would correlate to specific crime subtypes, thereby increasing the predictive performance for a particular type of criminal offender (Ormachea et al., 2017). With increased sample size, we will apply the predictive model to

criminal subtypes (e.g., arson, mass shooting) and hope to be able to draw connections between neurocognitive domains and more specific crimes. By examining criminal subtypes and identifying neurocognitive areas of interest, we hope to be able to better understand the drivers of crime, and also offer more effective and targeted rehabilitation direction for specific services.

The NCRA is able to predict reoffense at a level comparable to current risk assessments and has the additional benefits of being intuitive to use, easy to interpret, and uniformly deployable across age, gender, and crime level. Additionally, it holds the potential to give previously unavailable insights into the underlying neurocognitive drivers of decision making of criminal behavior.

DATA AVAILABILITY STATEMENT

The datasets for this manuscript are not publicly available because the data contain private health information for a protected participant class (prisoners and probationers). Requests to access the datasets should be directed to davideagleman@stanford.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Solutions IRB. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

GH, SD, and DE contributed to the conception and design of the study. GH, SD, PO, and DE contributed to the assessment invention. DW collected data for the study. GH performed all of the statistical analysis and machine learning. GH, SD, ES, PO, and DE wrote the manuscript.

FUNDING

The study was funded by the Mind Science Foundation.

ACKNOWLEDGMENTS

We would like to thank Brian Lovins, Luke Grove, Siji Brown, and Travis Pratt, countless assessors, and all the participants who volunteered at the Harris County Community Supervision and Corrections Department. We would also like to acknowledge Philip Abraham for his contribution on tuning parameters and machine learning.

REFERENCES

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The meta-analysis of clinical judgment project: fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist* 34, 341–382. doi: 10.1177/0011000005285875

Alper, M., Durose, M. R., and Markman, J. (2018). *Update on Prisoner Recidivism: A 9-Year Follow-up Period (2005-2014)*. Bureau of Justice Statistics Special Report, NCJ 250975. Washington, DC: U.S. Department of Justice.

Andrews, D., and Dowden, C. (2007). The risk–need–responsivity model of assessment and human service in prevention and corrections: crime-prevention jurisprudence. *Can. J. Criminol. Crim. Justice* 49, 439–464. doi: 10.3138/cjccj.49.4.439

Andrews, D. A., and Bonta, J. (1995). *The Level of Service Inventory - Revised*. Toronto: Multi-Health Systems.

Austin, J. (2004). The proper and improper use of risk assessment in corrections. *Fed. Sentenc. Rep.* 16, 194–199. doi: 10.1525/fsr.2004.16.3.194

Baron-Cohen, S., Jolliffe, T., Mortimore, C., and Robertson, M. (1997). Another advanced test of theory of mind: evidence from very high functioning adults with autism or asperger syndrome. *J. Child Psychol. Psychiatry* 38, 813–822. doi: 10.1111/j.1469-7610.1997.tb01599.x

Barry-Jester, A. M., Casselman, B., Goldstein, D., Conlen, M., Fischer-Baum, R., and Rossback, A. (2015). *Should Prison Sentences be Based on Crimes That Haven't Been Committed Yet?* Available at: https://fivethirtyeight.com/features/prison-reform-risk-assessment/ (accessed January 10, 2020).

Berk, R. (2017). An impact assessment of Machine Learning risk forecasts on parole board decisions and recidivism. *J. Exp. Criminol.* 13, 193–216. doi: 10.1007/s11292-017-9286-2

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2018). Fairness in criminal justice risk assessments: the state of the art. *Sociol. Methods Res.* doi: 10.1177/0049124118782533

Berk, R., and Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. *Fed. Sentenc. Rep.* 27, 222–228. doi: 10.1525/fsr.2015.27.4.222

Berk, R. A., and Bleich, J. (2013). Statistical procedures for forecasting criminal behavior. *Criminol. Public Policy* 12, 513–544. doi: 10.1111/1745-9133.12047

Bonta, J., and Andrews, D. A. (2007). Risk-need-responsivity model for offender assessment and rehabilitation. *Rehabilitation* 6, 1–22.

Campbell, M. A., French, S., and Gendreau, P. (2009). The prediction of violence in adult offenders. *Crim. Justice Behav.* 36, 567–590. doi: 10.1177/0093854809333610

Casselman, B., and Goldstein, D. (2015). *The New Science of Sentencing*. Available at: https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing (accessed August 14, 2019)

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Art. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Clémençon, S., and Vayatis, N. (2010). Overlaying classifiers: a practical approach to optimal scoring. *Constr. Approx.* 32, 619–648. doi: 10.1007/s00365-010-9084-9

Cortes, C., and Mohri, M. (2003). AUC optimization vs. error rate minimization. *Adv. Neural Inform. Proc. Syst.* 11, 356–360.

Desmarais, S., and Singh, J. (2013). *Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States*. Washington, D.C: Bureau of Justice Assistance.

Desmarias, S. L. (2013). Understanding risk assessments and its applications. *Paper Presented at the Justice and Mental Health Collaboration Program National Training and Technical Assistance Event*, Washington, DC.

Desmarais, S. L., Johnson, K. L., and Singh, J. P. (2016). Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychol. Serv.* 13, 206–222. doi: 10.1037/ser0000075

Desmarais, S. L., and Lowder, E. M. (2019). *Pretrial Risk Assessment Tools: A Primer for Judges, Prosecutors, and Defense Attorneys*. Available at: http://www.safetyandjusticechallenge.org/resource/pretrial-risk-assessment-tools-a-primer-for-judges-prosecutors-and-defense-attorneys/ (accessed October 9, 2019).

Domes, G., Hollerbach, P., Vohs, K., Mokros, A., and Habermeyer, E. (2013). Emotional empathy and psychopathy in offenders: an experimental study. *J. Personal. Disord.* 27, 67–84. doi: 10.1521/pedi.2013.27.1.67

Douglas, T., Pugh, J., Singh, I., Savulescu, J., and Fazel, S. (2017). Risk assessment tools in criminal justice and forensic psychiatry: the need for better data. *Eur. Psychiatry* 42, 134–137. doi: 10.1016/j.eurpsy.2016.12.009

Dressel, J., and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* 4:eaao5580. doi: 10.1126/sciadv.aao5580

Duwe, G. (2017). *Why Inter-Rater Reliability Matters for Recidivism Risk Assessment (Policy Brief Number 2017-03)*. Washington, DC: The Risk Assessment Clearinghouse.

Duwe, G., and Rocque, M. (2017). The effects of automating recidivism risk assessment on reliability, predictive validity, and return on investment (ROI). *Criminol. Public Policy* 16, 235–269. doi: 10.1111/1745-9133.12270

Eagleman, D. M. (2011). *Incognito: The Secret Lives of the Brain*. New York, NY: Pantheon Books.

Eaglin, J. M. (2017). Constructing recidivism risk. *Emory L. J.* 67, 59.

Eckhouse, L., Lum, K., Conti-Cook, C., and Ciccolini, J. (2019). Layers of bias: a unified approach for understanding problems with risk assessment. *Crim. Justice Behav.* 46, 185–209. doi: 10.1177/0093854818811379

Fazel, S., Singh, J. P., Doll, H., and Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis. *BMJ* 345:e4692. doi: 10.1136/bmj.e4692

Gottfredson, D. M. (1987). Prediction and classification in criminal justice decision making. *Crime Justice* 9, 1–20. doi: 10.1086/449130

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., and Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychol. Assess.* 12, 19–30. doi: 10.1037//1040-3590.12.1.19

Hanley, J. A., and Mcneil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

Harcourt, B. E. (2015). Risk as a proxy for race. *Fed. Sentenc. Rep.* 27, 237–243. doi: 10.1525/fsr.2015.27.4.237

Harris, K., and Paul, R. (2017). *Kamala Harris and Rand Paul: To Shrink Jails, Let's Reform Bail*. Available at: https://www.nytimes.com/2017/07/20/opinion/kamala-harris-and-rand-paul-lets-reform-bail.html (accessed October 9, 2019).

Holder, E. (2014). *Speech presented at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference, Philadelphia*. Available at: https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th (accessed August 8, 2019).

Howard, P. D. (2016). The effect of sample heterogeneity and risk categorization on area under the curve predictive validity metrics. *Crim. Justice Behav.* 44, 103–120. doi: 10.1177/0093854816678899

Kehl, D., Guo, P., and Kessler, S. (2017). *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*. Cambridge, MA: Harvard Law School.

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer.

Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkins, L., and Hansen, K. T. (2007). Actuarial assessment of risk for reoffense among adult sex offenders. *Crim. Justice Behav.* 34, 37–59. doi: 10.1177/0093854806291157

Lowenkamp, C. T., Holsinger, A. M., Brusman-Lovins, L., and Latessa, E. J. (2004). Assessing the inter-rater agreement of the level of service inventory revised. *Fed. Probation* 68, 34–38.

Markman, J. A., Durose, M. R., Rantala, R. R., and Tiedt, A. D. (2016). *Recidivism of Offenders Placed on Federal Community Supervision in 2005: Patterns from 2005 to 2010* (Washington, D.C: U.S. Department of Justice), 1–16.

Monahan, J., and Skeem, J. L. (2014). Risk redux. *Fed. Sentenc. Rep.* 26, 158–166. doi: 10.1525/fsr.2014.26.3.158

Ormachea, P. A., Davenport, S., Haarsma, G., Jarman, A., Henderson, H., and Eagleman, D. M. (2016). Enabling individualized criminal sentencing while reducing subjectivity: a tablet-based assessment of recidivism risk. *AMA J. Ethics* 18, 243–251. doi: 10.1001/journalofethics.2016.18.3.stas1-1603

Ormachea, P. A., Lovins, B. K., Eagleman, D. M., Davenport, S., Jarman, A., and Haarsma, G. (2017). The role of tablet-based psychological tasks in risk assessment. *Crim. Justice Behav.* 44, 993–1008. doi: 10.1177/0093854817714018

Osoba, O. A., Benjamin, B., Jessica, S., Luke Irwin, J., Mueller, P. A., and Cherney, S. (2019). *Algorithmic Equity: A Framework for Social Applications*. Santa Monica, CA: RAND Corporation.

Pope-Sussman, R., and Turner, S. (2015). Available at: https://www.courtinnovation.org/publications/strengths-and-limitations-risk-assessment-professor-susan-turner-university-california (accessed June 12, 2019).

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rice, M., and Harris, G. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law Hum. Behav.* 29, 615–620. doi: 10.1007/s10979-005-6832-7

Seidel, E. M., Pfabigan, D. M., Keckeis, K., Wucherer, A. M., Jahn, T., Lamm, C., et al. (2013). Empathic competencies in violent offenders. *Psychiatry Res.* 210, 1168–1175. doi: 10.1016/j.psychres.2013.08.027

Singh, J. P., Desmarais, S. L., and Van Dorn, R. A. (2013). Measurement of predictive validity in violence risk assessment studies: a second-order systematic review. *Behav. Sci. Law* 31, 55–73. doi: 10.1002/bsl.2053

Singh, J. P., Grann, M., and Fazel, S. (2011). A comparative study of violence risk assessment tools: a systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clin. Psychol. Rev.* 31, 499–513. doi: 10.1016/j.cpr.2010.11.009

Skeem, J. L., Monahan, J., and Lowenkamp, C. T. (2016). Gender, risk assessment, and sanctioning: the cost of treating women like men. *Law Hum. Behav.* 40, 580–593. doi: 10.1037/lhb0000206

Spohn, C. C. (2008). *How Do Judges Decide?: The Search for Fairness and Justice in Punishment*, 2nd Edn. Thousand Oaks, CA: Sage Publications.

Sreenivasan, S., Kirkish, P., Garrick, T., Weinberger, L. E., and Phenix, A. (2000). Actuarial risk assessment models: a review of critical issues related to violence and sex-offender recidivism assessments. *J. Am. Acad. Psychiatry Law* 28, 438–448.

State v. Loomis (2016). 881 N.W. 2d 749, 767 (Wis. 2016).

Tollenaar, N., and van der Heijden, P. G. (2019). Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. *PLoS One* 14:e0213245. doi: 10.1371/journal.pone.0213245

Tollenaar, N., and van der Heijden, P. G. M. (2013). Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *J. R. Stat. Soc. A* 176, 565–584. doi: 10.1111/j.1467-985x.2012.01056.x

Trevor, H., Robert, T., and Jerome, F. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin: Springer.

Ward, T., and Fortune, C. (2016). The role of dynamic risk factors in the explanation of offending. *Aggress. Violent Behav.* 29, 79–88. doi: 10.1016/j.avb.2016.06.007

Yang, M., Wong, S. C., and Coid, J. (2010). The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychol. Bull.* 136, 740–767. doi: 10.1037/a0020473

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., Conway, K. P., et al. (2014). NIH Toolbox Cognition Battery (CB): validation of executive function measures in adults. *J. Int. Neuropsychol. Soc.* 20, 620–629. doi: 10.1017/S1355617714000472

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

CONCEPTUAL ANALYSIS
published: 24 April 2020
doi: 10.3389/fpsyg.2020.00627

Adolescent Brain Development and Progressive Legal Responsibility in the Latin American Context

Ezequiel Mercurio[1], Eric García-López[2], Luz Anyela Morales-Quintero[3], Nicolás E. Llamas[4], José Ángel Marinaro[4] and José M. Muñoz[5]*

[1] Center of Interdisciplinary Forensic Research, Buenos Aires National Academy of Sciences, Buenos Aires, Argentina, [2] Instituto Nacional de Ciencias Penales, Mexico City, Mexico, [3] Criminology Program, Faculty of Law and Social Sciences, Benemérita Autonomous University of Puebla, Puebla, Mexico, [4] Department of Law and Political Science, National University of La Matanza, San Justo, Argentina, [5] Department of Psychology, Universidad Europea de Valencia, Valencia, Spain

In this article, we analyze the contributions of neuroscience to the development of the adolescent brain and shed additional light on the minimum age of criminal responsibility in the context of Latin America. In neurobiology, maturity is perceived to be complex because the brain's temporal development process is not uniform across all its regions. This has important consequences for adolescents' behavior; in their search for the acceptance of their peers, they are more vulnerable to pressure and more sensitive to stress than adults. Their affectivity is more unstable, and they show signs of low tolerance to frustration and important emotional reactivity, with a decrease in the capacity to self-regulate. Consequently, risky behavior presents itself more frequently during adolescence, and behaviors that transgress norms and social conventions typically peak between the ages of 17 and 19 years. However, only a small percentage of young offenders escalate their behavior to committing crimes during adulthood. In comparative law, there are considerable differences in Latin American countries' legal dispositions regarding the minimum age of criminal responsibility; Brazil, Costa Rica, and Ecuador regard the age of criminal responsibility to be 12 years, while Argentina accepts this to be 16 years. From a legal viewpoint, however, the debate about the minimum age of criminal responsibility is connected to other circumstances that, because they are still at a developmental stage, are attributed to adolescents' rights in their decision-making and understanding of autonomy (e.g., the minimum ages for voting, alcohol consumption, and medical consent). We argue that research on the development of the adolescent brain does not provide definitive answers about the exact age required for different juridical purposes. Nonetheless, the current state of knowledge does allow for reflection on the development and maturation of adolescents and the implications for considering them criminally responsible. It also validates demands for a system that provides adolescents with greater protection and that favors their healthy integral development. In any case, although a specific minimum age is not evident, this study is disposed not to recommend lowering the age of criminal responsibility, but rather increasing it.

Keywords: neurolaw, juvenile criminal behavior, juvenile criminal law, adolescent brain, legal responsibility

OPEN ACCESS

Edited by:
Daniela Smirni,
University of Palermo, Italy

Reviewed by:
Maria Ioannou,
University of Huddersfield, United Kingdom
Colleen Berryessa,
Rutgers, The State University of New Jersey, United States

****Correspondence:***
Eric García-López
garcialopez@gmx.com

Specialty section:
This article was submitted to Neuropsychology, a section of the journal Frontiers in Psychology

Received: *21 December 2019*
Accepted: *16 March 2020*
Published: *24 April 2020*

Citation:
Mercurio E, García-López E, Morales-Quintero LA, Llamas NE, Marinaro JÁ and Muñoz JM (2020) Adolescent Brain Development and Progressive Legal Responsibility in the Latin American Context. Front. Psychol. 11:627. doi: 10.3389/fpsyg.2020.00627

INTRODUCTION

Studies of human development often define adolescence as a complex transitional phase between childhood and adulthood. However, from a neuroscientific point of view, it is not easy to define or delimit this age group; if we take into account the fact that cognitive abilities and different brain regions do not develop uniformly or simultaneously, this is even more the case. Moreover, the complex processes of brain and cognitive development are intimately influenced by culture and environment.

This complexity has also influenced the law, as is evidenced by the variety of legislation, which considers differences in the minimum age of responsibility according to various types of activities or decisions and which could also have civil and penal consequences in adulthood.

The difficulty in defining adolescence has recently been identified by the United Nations Committe on the Rights of the Child [CRC] (2016), in General Comment No. 20. This observation centers on the temporal concept of childhood, "from 10 years until the 18th birthday" (United Nations Committe on the Rights of the Child [CRC], 2016, para. 5). In the same manner, it gives perspective to the complexity of this definition, which, among other reasons, lies in the difficulty of identifying an exact age from a biological point of view. In particular, the observation notes that "different brain functions mature at different times" (United Nations Committe on the Rights of the Child [CRC], 2016, para. 5).

Following this line of thought, we can observe a great interest in the development of the adolescent brain, which forms the focus of a variety of studies that specialize in adolescence and human rights. Such is the case with UNICEF's recent reports on adolescence in Argentina (UNICEF, 2017) and on lowering the age of criminal responsibility in Uruguay (UNICEF, 2014). At the same time, research efforts have been strengthened by projects such as the Adolescent Brain Cognitive Development Study[1] (Feldstein Ewing et al., 2018), which follows 10,000 children between the ages of 9 and 10 years over the course of a decade, utilizing neuroimaging studies, neuropsychological evaluations, and various non-specific health investigations.

This area of knowledge, together with the development of modern neuroimaging techniques, has begun to influence different legal systems, particularly those of the Anglosphere tradition, where neuroscientific arguments have been presented in different penal cases during the last decade (Farahany, 2015; Altimus, 2017). Specifically, knowledge of how the brain grows, matures, and develops during adolescence, and its relationship with behavior, has started to influence the law (Mercurio, 2012; Steinberg, 2013; Cohen and Casey, 2014; Jones et al., 2014). For example, the Supreme Court of the United States has used arguments based on neuroscience to inform decisions about penal cases in which adolescents have been involved; this can be seen in the cases Roper v. Simmons (2005), Graham v. Florida (2010), and Miller v. Alabama (2012).

In this article, we analyze the contributions of neuroscience to knowledge of the development of the adolescent brain and shed additional light on the minimum age of criminal responsibility in the Latin American context.

THE BRAIN AND ADOLESCENCE

It is evident that young people and adolescents are different from adults. Research articles in the field of neuroscience have shown that it is possible to ascertain, in terms of neurobiology, the reasons for these differences. The growth and development of the brain obey the interaction between genetics and the environment (nature and nurture), modeled by the characteristics of the different evolutionary stages of human development. While in the prenatal stage genes play a key role in the formation of the different brain circuits, during the stages following birth it is experiences and interaction with the environment that influences these circuits (Pascual Urzúa, 2014).

The complex demands of the environment during development require the modification of the brain's connections. On this subject, Churchland (2012) indicates that human beings are born with immature brains and that this is actually an evolutionary advantage as it makes it possible to obtain a greater benefit from interactions with the environment, while also allowing for adaptation to the complex physical and social world. Synaptic connections (formed through synaptogenesis) are modified during different evolutionary stages, and depending on the region of the brain, they reach their maximum expansion between 2 and 7 years of age. This is followed by a process of the elimination of connections (synaptic pruning), which is widely accepted to last until the end of adolescence in the prefrontal region. There is concrete evidence to show that synaptic pruning in the prefrontal cortex also occurs between 20 and 30 years of age (Petanjek et al., 2011). Thus, it can be said that the process of expansion occurs during childhood, whereas the process of contraction and the elimination of connections occurs during adolescence and beyond and is followed by stabilization during adulthood (Giedd et al., 1999; Gogtay et al., 2004; Pascual Urzúa, 2014). The actual hypothesis about this process is that the large neuronal expansion of the connections during childhood allows children to have a broad connection with their physical, cultural, and social environment. After this time, the most requested and strengthened connections will prevail, whereas those that are less needed will be eliminated (Pascual Urzúa, 2014).

In neurobiology, maturity is perceived to be complex because the brain's temporal development process is not uniform across all its regions. Regions related to sensory and motor activities show a pattern unlike those related to cognitive and complex affective functions, such as the executive functions (these functions are called "the most human functions of men" by Luria) or those related to the socioemotional process (e.g., empathy). In this sense, recent studies have specifically shown that the frontal lobe finishes maturing at ˜30 years of age, later than the other regions (Østby et al., 2009; Tamnes et al., 2010; Petanjek et al., 2011). This has important consequences for adolescents' behavior.

On this point, Dahl (2004) notes the existence of a paradox in adolescents' health; while we see increased physical growth, the strengthening of the immune system, and overall better cognitive

[1] For further information, refer to http://abcdstudy.org.

abilities compared to childhood, morbidity and mortality increase by 200% over the same period. This is connected to the difficulties associated with adolescents having to control their behavior and manage their emotions (Kelley et al., 2004; García-López and Mercurio, 2019), which makes them more vulnerable to risky behavior (Steinberg, 2004; Gardner and Steinberg, 2005; Barbalat et al., 2009; Pfeifer et al., 2011). Clear examples of this are reckless driving, alcohol and drug consumption, and violence, while such behavior may also lead to accidents, suicide, depression, eating disorders, and risky sexual behaviors (Dahl, 2004; Eaton et al., 2008). Adolescence is therefore seen as a period of great opportunity but, at the same time, great vulnerability (García-López, 2004).

This paradox is understandable because of scientific evidence from studies about the relationship between brain development and the manifestation of risk behaviors in adolescence. This is explained in the following subsections.

Maturity Gap

As we noted above, during adolescence, the brain and cognitive abilities do not develop at the same rate. Regions that seek reward are more active and mature earlier than the regions controlling impulses. This model is known as the "dual system" or "maturational imbalance model" (Casey et al., 2008; Steinberg, 2008). Based on analyses of more than 900 individuals between the ages of 10 and 30 years, Steinberg et al. (2009a) observed that cognitive capacity, for example, logical reasoning and memory, matures by age 16 years. Nevertheless, psychosocial maturity – self-control and future orientation, especially in the presence of peers and social contexts – does not fully mature until the person is in his/her 20s. In a large sample (N = 5,404), which contained individuals between the ages of 10 and 30 years in 11 countries, Icenogle et al. (2019) found that during adulthood, individuals' sensation seeking declined, and their impulse control, future orientation, and resistance to peer influence increased. This study suggests adolescents achieve the same cognitive abilities as adults at age 16 years, but their psychosocial maturity is not developed until their 20s. These results are similar to those of previous studies (Steinberg et al., 2009a,b; Chein et al., 2011; Quinn and Harden, 2013; Shulman et al., 2014). This "maturity gap" between cognitive and psychosocial development is a window of opportunity to increase the chances of making risky decisions, leading to risky behaviors during adolescence.

Reward-Based Behavior

The evidence about the behavior incentive process has allowed researchers to identify different brain circuitry and the important role of dopamine in these circuits. In accordance with what was explained in the subsection discussing the maturity gap during adolescence, an imbalance has been found between the reward and the regulatory circuitry.

This imbalance explains the increased reward-seeking behavior in this period, which includes monetary, novel, and social rewards, as well as the dopamine system's sensitivity to rewards (Galvan, 2010).

Considering Galvan's review, the dopamine system is hyperresponsive or overcommitted in its response to rewards during adolescence. This increases the tendency to seek novelty and sensations (Dahl, 2004).

Several studies have found an important availability and function of dopamine during adolescence, which can be explained by the dopamine system's high sensitivity to reward, the search for reward, and sensation-seeking behaviors (Luna et al., 2013). In addition, this situation increases the susceptibility of adolescents to the motivational properties of substance abuse (Casey and Jones, 2010). Studies, such as the one by Casey et al. (2008), have found that risk behaviors have a neurobiological correlation with the responses to rewards. Specifically, they have found that adolescents who have sexual escapades, drink excessively, practice high-impact sports, and engage in similar activities show greater activity in the nucleus accumbens–frontal cortex, especially when they play to earn money.

Luna et al. (2013) investigated the relationship between rewards and inhibitory control using incentives based on task performance. In this study, the younger participants, children and adolescents, showed more problems with inhibiting their responses compared to adults. The adolescents took longer to complete the task, but they showed high activation in the brain regions of the reward system. This supports the idea that these behaviors lead to acquiring some reward.

In another study, Palminteri et al. (2016) compared how both adults and adolescents learn to make choices based on the information they have available. The results of their research showed that teenagers focus on rewards and find it difficult to learn to avoid punishment or consider the consequences of their actions. The volunteers had to choose symbols associated with either a reward or punishment or a symbol without a consequence. After the choice was made in each task, the participants received feedback on their performance. Adults learned faster from their experience and modified their responses. They also avoided the symbols associated with punishment and learned from the feedback to make better decisions, while teenagers had more trouble doing so.

It is important to note that the evidence about reward-based behavior is most evident in contexts of heightened arousal. For example, McKewen et al. (2019) found that adolescents who had greater behavioral arousal during a task in which they argued with their parents showed a lower regulatory ability with a lower heart rate variability rate than those who had low behavioral arousal.

Peer Pressure and Reward Sensitivity

During adolescence, peer pressure plays a key role in behavior (Currie et al., 2004; Prinstein et al., 2011). Adolescents engage in riskier behaviors when they are with their peers than when they are alone. This is known as the "peer effect" (Gardner and Steinberg, 2005; Albert and Steinberg, 2011; Albert et al., 2013; Smith et al., 2014). The "peer effect" and reward sensitivity are interrelated and have a powerful influence on adolescent risk taking. In adolescence, a social context increases activity in reward brain regions and leads to changes in the processing of rewards, which leads to risky behaviors (Chein et al., 2011; Ciranka and van den Bos, 2019). In an experimental study, Gardner and Steinberg (2005) reported that early and late adolescents took more risks on a computerized driving task

when they were with their peers, although adults showed no difference in the amount of risky driving related to the social context. Somerville et al. (2011) found that adolescents are particularly sensitive to the reward-sensitizing effects of social stimuli, but they stated that this sensitization may affect their inhibitory control. Chein et al. (2011) employed functional magnetic resonance imaging (fMRI) during a video driving task and suggested that, in the presence of peers, adolescents had increased activation in their reward brain regions and evidenced higher levels of risky driving. In a recent article, Smith et al. (2018) used probabilistic gambling and go/no-go tasks while 28 adolescents (aged 15–17 years) were in an fMRI. They found an activation of the striatum and anterior insula when adolescents made risky decisions in the presence of peers, but this presence "had minimal impact on the engagement of typical cognitive control regions." The authors state that these results support the conclusion that when adolescents are with their peers they recruit reward-processing regions. This increases their reward sensitivity and thus leads to risky decision-making, although their capacity to engage in self-control does not diminish.

Different studies have highlighted the importance of peers and peer groups in the initiation of alcohol and drug consumption (Spear, 2000; Dishion and Tipsord, 2011; Trucco et al., 2011). In their search for the acceptance of their peers, adolescents are more vulnerable to pressure and more sensitive to stress than adults. Their affectivity is more unstable, and they show signs of a low tolerance for frustration and important emotional reactivity with a decrease in their capacity for self-regulation. These characteristics affirm that adolescents lack the same level of emotional, cognitive, or behavioral maturity as adults. Adolescents make decisions differently than mature people (Kambam and Thompson, 2009), and they overestimate short-term benefits.

Using a rodent model, Logue et al. (2014) found that juvenile mice, but not adults, increased their consumption of alcohol when their peers were present. These results suggest that during adolescence the presence of peers increases reward sensitivity, and this mechanism has been conserved among mammalian species (Trezza et al., 2011; Logue et al., 2014).

Risky Decision-Making

During adolescence, there is an important increase in behaviors that transgress norms and social conventions, peaking between the ages of 17 and 19 years (Federal Bureau of Investigation, 2003; Loeber et al., 2011). In general, adolescents commit antisocial behaviors in peer groups, and adults do so alone (Albert et al., 2013). However, only a small percentage of young offenders escalate their behavior to committing crimes during adulthood (Loeber et al., 2011).

Why does risky behavior present itself more frequently during adolescence? At present, there is important scientific evidence showing that frontal brain regions, which are related to organization, planning, and inhibitory control, are not fully developed until the end of adolescence (the third decade of life), and these regions are the last to mature (Spear, 2000; Galvan et al., 2006; Tamnes et al., 2010; Spear, 2013; Hartley and Somerville, 2015). On the other hand, regions that are reward-sensitive and regions connected to emotions are shown to be more active (Spear, 2000; Sowell et al., 2004; Toga et al., 2006; Giedd, 2008; Hartley and Somerville, 2015) and to have greater emotional reactivity (Guyer et al., 2016). This greater activity could be related to a sensitivity to focusing on possible gains in the short term, despite the negative consequences this might bring in the future. This way, the temporal distance between the maturation of both rational and emotional systems and their fragile communication generate a period of high vulnerability to risky behavior (Steinberg et al., 2009a; Icenogle et al., 2019). Different articles (Casey et al., 2008; Steinberg, 2008; Luna and Wright, 2016) know this framework as "dual systems" or "maturational imbalance." This model states that the difference in the development of sensation-seeking behaviors and self-control leads to a preference for behaviors that seek reward, novelty, and risk (Smith et al., 2018).

Exposure to risky behaviors (such as unprotected sexual intercourse, the consumption of toxic substances, or, most critically, the antisocial behaviors that tend to occur with greater intensity during adolescence) indicates that adolescents have less behavioral capability to prevent damage, despite the presence of more developed cognitive abilities. How is this possible? Can adolescents know the theoretical consequences of their actions but still fail to effectively inhibit them? The answer is related to the interaction among environmental factors, cerebral immaturity, and a marked decrease in activity in the prefrontal regions and their neural connections. There is also a smaller response to aversive stimuli and an increase in activity registered in regions related to the reward circuit and emotional reactivity.

Not only do structural and functional modifications in the prefrontal region improve self-control, but they also improve connections in areas related to emotions, such as the limbic system, allowing for an improvement in the interaction between cognition and emotions (Steinberg, 2008). Past studies reported the areas that regulate the processing of rewards, social information, and emotions are more sensitive and more easily aroused around middle adolescence (Giedd, 2004, 2008; Blakemore and Choudhury, 2006; Poon, 2018). This effective coordination between cortical and subcortical regions and the cognition–emotion interface encourage the modulation of activations sparked by social and affective stimuli, thus allowing deliberate reasoning. Likewise, this process is bidirectional, modulating the excessively deliberate decision-making with social and emotional information (Steinberg, 2008). As noted by Steinberg (2008), these modifications put a stop to the impulsive search for sensations and give a greater resistance to peer influence. These two factors together should decrease risk-taking; this usually occurs during adulthood.

THE MINIMUM AGE OF LEGAL RESPONSIBILITY IN LATIN AMERICA

The described conditions, common of development during adolescence, have been acknowledged in the international regulatory framework through different judicial documents. For

example, the Convention on the Rights of the Child (United Nations General Assembly [UNGA], 1989) indicates that "the child, by reason of his physical and mental immaturity, needs special safeguards and care" and promulgates the importance of contributing to the enforcement of children's rights to survival and healthy development.

Furthermore, Principle 2 of the Declaration of the Rights of the Child states that:

> The child shall enjoy special protection, and shall be given opportunities and facilities, by law and by other means, to enable him to develop physically, mentally, morally, spiritually and socially in a healthy and normal manner and in conditions of freedom and dignity (United Nations General Assembly [UNGA], 1959).

Similarly, Article 3 of the Convention on the Rights of the Child states that every measure taken by any public or private institution of social welfare, courthouse, government authority, or regulatory body that relates to children must consider the best interests of the child. The same criterion is used for any general commentaries made by the UN Committee on the Rights of the Child and for the advisory opinions made in the Inter-American Court of Human Rights.

In the preamble of the Convention on the Rights of the Child, it is established that children require "special care," and in Article 19 of the American Convention on Human Rights (Organization of American States [OAS], 1969), it is indicated that they ought to receive "measures of protection." In the case of children responsible for committing crimes, the universal system of human rights has developed the United Nations Standard Minimum Rules for the Administration of Juvenile Justice (also known as the Beijing Rules), the United Nations Guidelines for the Prevention of Juvenile Delinquency (The Riyadh Guidelines), and the United Nations Rules for the Protection of Juveniles Deprived of their Liberty. All of these highlight the need to adopt measures of specific care for minors, always taking into account their vulnerable situation as a result of their immaturity, inexperience, and mid-development status.

However, from a legal point of view, full agreement regarding the ages at which the concept of childhood is applicable has yet to be reached. For example, the Convention on the Rights of the Child establishes in Article 1 that: "For the purposes of the present Convention, a child means every human being below the age of 18 years unless under the law applicable to the child, majority is attained earlier." In this sense, international regulations establish the importance of different treatment for children and adults, suggesting 18 years old as an appropriate age to make the distinction, but offering flexibility regarding the establishment of the age of jurisdiction according to different countries' legislation. There is general agreement on the difference between children and adults but not on the age range that distinguishes one from the other. In paragraph 3 of Article 40 of the Convention, it is asserted that participating states must try to promote, among other things, the establishment of a minimum age, so that children under that age can be presumed not to have the capacity to disobey the penal law, but a concrete minimum age is not mentioned.

From a legal viewpoint, the debate about the minimum age of criminal responsibility is connected to other circumstances that, because they are still at a developmental stage, are attributed to adolescents' rights in their decision-making and understanding of autonomy, such as the minimum ages for voting, buying cigarettes, consuming alcohol, medical consent, and accessing contraception. It is as if, on the one hand, adolescents' capacity to make decisions and to take responsibility for their own actions is recognized, while, on the other hand, when convenient, this is not acknowledged.

This is evident in the application of certain public policies where, for example, the legal smoking and drinking age is established at 18 years old and the legal age for accessing contraception is 14 years. In that sense, countries like the United States set a high drinking age – 21 years old – while, as shown by a recent report by the American Academy of Pediatrics, the range for the legal smoking age is recommended to be between 18 and 21 years old[2].

These differences can also be seen in different Latin American countries, as shown in **Table 1**.

For example, the Argentine Civil and Commercial Code (CCyC) (Ministry of Justice and Human Rights Argentina, 2014) holds the definition of child to be those who have not yet turned 13 years old and that of adolescent to be youngsters who are between 13 and 18 years of age (CCyC, art. 25); it also includes the concept of "progressive capabilities" (CCyC, art. 117). For decisions relating to health, it states that adolescents between the ages of 13 and 16 years can decide for themselves when it comes to health treatments that are either non-invasive or that present no risk to their health or lives (CCyC, art. 26). Adolescents older than 16 years are considered to be adults for decisions that relate to the care of their bodies (CCyC, art. 26). From the age of 13 upward, even if their parents oppose, adolescents can file a lawsuit if they have judicial authorization and provided that they have legal assistance during the process (CCyC, art. 678). They can also acknowledge paternity (CCyC, art. 680).

The Argentine Civil and Commercial Code establishes an interesting distinction between adolescents when it comes to types of decision-making. As shown in different publications, the same age does not necessarily mean the same capacity to perform all the acts of civic life. Progressive capacities show that while a 14-year-old adolescent has the competence to request contraception, he/she does not have it to consent to a surgical intervention (Herrera et al., 2015). In the same way, the voting age is 16 years, while the drinking and smoking age is 18 years, and the differences in the driving age depend on whether it is for motorbikes, cars, or public transportation (16, 17, and 21 years of age, respectively).

But how is it possible to reconcile the fact that adolescents are mature enough to, for example, ask for contraceptives while being younger than 18 years, or to consent to surgery and vote

[2]https://www.aap.org/en-us/about-the-aap/aap-press-room/aap-press-room-media-center/Pages/Tobacco-and-E-Cigarettes.aspx (accessed January 10, 2020).

TABLE 1 | Minimum ages to exercise certain rights or to consume certain substances.

Country	Criminal age (age range)	Age of majority (and voting age)	Drinking age	Smoking age
Argentina	16–18	18 Voting at 16	18	18
Belize	Data not available	Data not available	18	18
Bolivia	14–18	18	18	18
Brazil	12–18	18 Voting at 16	18	18
Chile	14–18	18	18	18
Colombia	14–18	18	18	18
Costa Rica	12–18	18	18	18
Cuba	16–18	16	Data not available	Data not available
Dominican Republic	13–18	18	18	18
Ecuador	12–18	18 Voting at 16	18	18
El Salvador	12–18	18	18	18
Guatemala	13–18	18	18	18
Honduras	12–18	21	18	21
Mexico	12–18	18	18	18
Nicaragua	13–18	18 Voting at 16	18	18
Panama	12–18	18	18	18
Paraguay	14–18	18	18	18
Peru	14–18	18	18	18
Uruguay	13–18	18	18	18
Venezuela	14–18	18	18	Data not available

Source: Our own elaboration based on data found in Sedletzki (2016). Regarding data about minimum ages and voting ages, current civil codes from each country have been revised. The source consulted for the drinking ages was ICAP (2012). Specific regulations from each country have been revised for smoking age.

at 16 years, but not to smoke until 18 years or – in the area of criminal law – to be punished if they are below the age of 16 years?

In this regard, knowledge based on neuroscience explains the fact that the decision-making process depends on the type of decision, the environment, and the context in which adolescents find themselves. In other words, it can be asserted that adolescents are mature enough to make certain decisions in determinate circumstances, but not to make others.

This debate arose in the United States as a result of two cases that reached the Supreme Court; they centered on reports made by the American Psychological Association (APA), also known as the APA (Steinberg et al., 2009a). In the first case, Hodgson v. Minnesota (1990), the discussion was over an adolescent's right to interrupt her pregnancy without previously notifying both of her parents. The APA argued that, taking into account the scientific evidence available, adolescents between the ages of 14 and 15 years showed no differences compared to adults, either in quality or in quantity, regarding logical reasoning in the comprehension of medically informed decisions (American Psychological Association, 1990). That is to say, it maintained the criterion of adolescents' maturity to make medically informed decisions.

In Roper v. Simmons (2005), the death penalty for adolescents was abolished, and the American Psychological Association (2005) asserted that the immaturity that leads to the lesser culpability of adolescents is grounded in three aspects:

1. a lack of development of the sense of responsibility, which makes it difficult to control impulses;
2. a high vulnerability to peer pressure;
3. adolescents' personality not yet being completely formed, causing their personality traits to be more transitory than fixed.

In this case, the APA continued to support the criterion that asserts adolescents' immaturity as a reason to not convict them as if they were adults.

This apparent contradiction was highlighted in the Roper case, to which the APA responded by pointing out that both cases dealt with very different issues; the first regarded adolescents' competence to consent to medically informed treatments, whereas the second related to adolescents' culpability in criminal law, and whether they can be convicted in the same manner as adults.

As previously mentioned, in the last few years there has been an increase in evidence of the different ages at which cognitive and psychosocial abilities develop and mature in adolescents; these abilities evolve and mature in different ways. That is to say, there is a temporal gap between the development of the cognitive abilities for information processing, the prefrontal cortex, which is mostly matured by the age of 16 years, and the development of the abilities that are required for coordination between affection and cognition – cortical and subcortical connections – the

maturation of which is completed at a later time (Steinberg, 2008; Icenogle et al., 2019).

The performance of intellectual and cognitive abilities will therefore not show significant improvement beyond the age of 16 years (Steinberg et al., 2009a). Meanwhile, psychosocial maturity, which is related to impulsivity, risk perception, sensation seeking, future orientation, and resistance to peer pressure, requires an effective coordination between emotions and cognition, and this occurs from the age of 20 years onward (Steinberg, 2008; Steinberg et al., 2009a). In neurobiological terms, cognitive tasks that require adequate interaction and coordination between multiple brain regions reach their development and maturity after the age of 16 years (Steinberg, 2009; Icenogle et al., 2019).

The improvement in connectivity between cortical and subcortical areas is related to the modification of susceptibility to peer pressure, which also influences risk-taking (Steinberg, 2008). Adolescents show socioemotional network activation when in the presence of their peers; this activation brings with it a decrease in self-control regulation and a greater exposure to risky behavior. This mechanism, in which peer pressure brings a greater exposure to risk-taking, occurs in the period between the ages of 19 and 20 years (Gardner and Steinberg, 2005; Steinberg and Monahan, 2007). Therefore, behavior in adolescents will differ depending on whether they are alone or with company, or if they are emotionally activated. In early adolescence, if the socioemotional circuit is not activated – for example, when adolescents are alone or in the company of an adult – there is bound to be greater cognitive control, which allows them to avoid exposure to risky situations. However, if they are accompanied by peers, or under certain conditions such as emotional activation, the socioemotional circuit is activated, which lowers their effective regulation of cognitive control. During adolescence, these circuits of cognitive control mature in such a way that, even though high socioemotional activation conditions may still be experienced during adulthood, inclinations toward risky behavior can be modulated (Steinberg, 2008).

Following this chain of ideas, in contexts where adolescents are not emotionally activated and do have time to make a decision, meaning they are "cold thinking," although cognitive control is still in development, it is sufficient to control impulses and promote more deliberate actions (Botdorf et al., 2017). Under these conditions, risk-taking is also like that of adults; informed medical decision-making and voting come under this context. On the other hand, in contexts where adolescents are emotionally activated, or when they are with their peers and do not have time to make a decision, meaning they are "hot thinking," adolescents find themselves in risky situations more frequently than adults (Burnett et al., 2010; Paulsen et al., 2011). Poon (2018) found a bell-shaped development curve in hot executive functions during adolescence with a peak at the ages of 14 and 15 years. The author stated that the sensitivity to reward and the risk-taking propensity were highest during this period.

In making decisions related to health, it is not only possible to consult with different doctors, but also with other specialists or parents; generally, medical decisions are not made under strict time constraints. These are the arguments put forward by the APA in the Hodgson case; when an adolescent contemplates the option of interrupting a pregnancy, she is taking time to think about her decision. During that time, she can consult with people she trusts or with different professionals (Steinberg et al., 2009a).

Some authors extend these arguments to other judicial contexts, such as the capacity to be on trial (Grisso et al., 2003), pointing out that the abilities that a person needs to be able to be tried include an understanding of the different stages of the process, the roles of each of the actors, and the meaning of the allegations, along with the ability to reason this information. They argue that, when it comes to these abilities, differences exist between adults and adolescents younger than 15 years, but not adolescents of 16 years of age.

In "hot thinking" contexts where adolescents are under pressure from their peers, under stress, and without adult supervision, the decisions they make and their behaviors are risky and reckless (Botdorf et al., 2017). In these contexts, they are less influenced by their theoretical knowledge about potentially negative consequences and so are more willing to take risks to potentially obtain short-term rewards (Hartley and Somerville, 2015). As previously shown, when adolescents are around their peers, their behavior becomes more impulsive, and the decisions they make become riskier (Hartley and Somerville, 2015). Smith et al. (2014) examined the influence of peers on adolescent risk-taking under a gambling task and found "that the presence of peers increases risky decision-making during adolescence even when explicit information about the probability of negative outcomes is provided, and even (perhaps especially) when these negative outcomes are portrayed as highly likely." These results suggest that when adolescents are in the presence of peers, providing adolescents with information about the likelihood of negative outcomes may not be as effective as expected.

FURTHER REMARKS ON LATIN AMERICAN LEGISLATION

The difficulty in the consideration of the legal responsibility of adolescents is evident when we look at cases in different countries. For example, in the case of the United States, Farahany explains:

> In a triology [*sic*] of cases [i.e., *Roper v. Simmons*, *Graham v. Florida*, and *Miller v. Alabama*], the United States Supreme Court has cited to evidence about the developing juvenile brain to find it unconstitutional under the Eighth Amendment of the United States Constitution to executive juveniles, to impose life without the possibility of parole for non-homicidal offenders, or to have a mandatory scheme of life imprisonment without the possibility of parole. Since the latest of these cases, *Miller v. Alabama*, there is considerable confusion and debate by lower courts about the meaning of that ruling and the extent to which a judge must consider neuroscience when sentencing a juvenile offender (Farahany, 2015).

Regarding the United Kingdom, Catley and Claydon (2015) state that it is "unlikely that neuroscientific advances in

understanding the brains of adolescents relevant to the age of criminal responsibility would appear in English case law." The Netherlands is another interesting case:

> The measure of "Placement in an Institution for Juveniles" ("Plaatsing in Inrichting Jeugdigen," PIJ, art 77s Criminal Code) can be imposed by the court for 3 years, and can thereafter be continued by the court to a maximum of 7 years. PIJ is intended for criminal juveniles with a developmental disorder or psychological/psychiatric problems. The aim of the PIJ-measure is reintegration into society by resocialization. In the Netherlands, juveniles of 12–18 years in principle fall under juvenile criminal law. Juveniles of 16 or 17 may be sentenced according to adult criminal law. Since the new "Adolescent Criminal Law" came into effect, Apr. 1, 2014, adolescents of 18–23 years old may be sentenced according to juvenile criminal law (de Kogel and Westgeest, 2015).

In Latin America, there are numerous human rights treaties that have been ratified by the different states and that govern this matter. With this in mind, and in consonance with Article 24 of the American Convention on Human Rights, the principle of equality must be understood to be the obligation to treat equals in the same way. It also means, however, that those not under equal conditions must not receive the same legal treatment. This is one of the reasons why children, adolescents, and adults should not be treated in the same way: as has already been explained, their cognitive abilities are not the same.

That said, rights and obligations must be implemented according to their context and the consequences that they carry with them. From this point of view, the objective of the Argentinean legislation mentioned earlier, which holds that those exercises of rights that might imply a long-term consequence for children and adolescents are the last rights to be acquired, appears appropriate. If these types of decisions were made in the context of peer pressure, or any other context of "hot thinking," it could bring legal consequences for those in this age group. Legal limitations that demand consent from the responsible adult, or a judicial decision (if the former does not give consent), allow for the protection of adolescents' integrity and development and force them to deliberately consider or debate their decisions. At the same time, the adolescent is treated as a "subject of rights" (and not "object of rights"): if their decision is not unreasonable or does not put them into a risky situation, they can do as they will.

This does not generate conflict as long as we are referring to the exercising of rights (such as the right to vote or to have control over one's own body) that carry inherent responsibilities. However, when we enter the realm of legal responsibilities, there are bigger differences in the legal dispositions in comparative law: Brazil, Costa Rica, and Ecuador regard 12 as the age of criminal responsibility, whereas Argentina views this age to be 16 years.

As has been noted, adolescents' brain development is not linear, and therefore it is not (yet) possible, from a neuroscientific perspective, to define the exact moment from which a person can act with absolute cognitive capacity (or at least a capacity appropriate to criminal responsibility). While this is true, it does not detract from the fact that recent studies have indicated that the development of the brain's executive functions is completed after the age of 21 years. Legislative debates on increasing the age of criminal responsibility are therefore needed, so that a person between the ages of 18 and 21 years will not receive the same treatment as an older adult, and so that they will not be seen as being over the minimum age of criminal responsibility.

As such, allowing a 12-year-old child to potentially be considered as criminally responsible presents a clear contradiction to the neuroscientific advances that have been made in recent decades. At the same time, this also constitutes a violation of the principle of equality as a 12-year-old child cannot receive the same legal treatment as a 16-year-old, because they are at different stages of cognitive development.

It would be wrong, however, to consider the determination of the minimum age of criminal responsibility to be the only relation between neuroscientific advances and juvenile criminal law. The increased cognitive development, the comparative decrease in the executive functions, the greater weight of peer pressure, and the underestimation of risk must also directly influence the principle of culpability and, consequently, the criminal response that an adolescent who is considered to be criminally responsible for a crime receives.

Taking this into consideration, in Latin American comparative law, it can be observed that a wide variety of socioeducational measures are considered as appropriate criminal consequences, including admonition, fines, community service, the obligation to finish schooling, apologies to victims, damage repair, and rehabilitation, among others (e.g., Chile: art. 6 from Law 20084; Colombia: art. 177 from Law 1098; Costa Rica: art. 121 from Law 7576; Ecuador: art. 378 from Law 100; Guatemala: art. 238 from Decree 27/2003; Honduras: art. 195 from Decree 73/96).

In this sense, it is necessary to highlight Law 287 of Nicaragua. In Article 95, it is stated that a person who was between the ages of 13 and 15 years at the moment of action and who was found to be criminally responsible for committing a crime will be sentenced with the application of socioeducational measures that do not involve the deprivation of freedom, whereas those who are 12 years of age or under are exempted from any criminal responsibility. At the same time, it imposes a maximum penalty of 6 years' imprisonment for adolescents between the ages of 16 and 18 years who are criminally convicted. This legislation is in harmony not only with the supranational legislation of human rights, but also with advances in neuroscience. Indeed, through legislation of this kind, the link between the gradual increase of criminal responsibility and the development of the adolescent brain can be demonstrated. It can therefore be cited as a very good example. On the contrary, Argentinian juvenile criminal law is considered to be incompatible with the region's current human rights treaties[3].

[3]For more on this, see Inter-American Court of Human Rights, *Mendoza et al. v. Argentina* (Preliminary Objections, Merits, and Reparations), May 14, 2013 (ser. C) No. 260, para. 295.

DISCUSSION

In recent decades (1984–2017), interest in the applications of neuroscience to law, and particularly criminal law, has increased notably (Farahany, 2015). For example, in juvenile criminal law, research on the maturation, growth, and development of the adolescent brain has had a big impact on the decisions taken by the Supreme Court of the United States (Graham v. Florida, 2010; Jackson v. Hobbs, 2012; Mercurio, 2012, 2014; Miller v. Alabama, 2012; Escobar et al., 2014).

There is much scientific evidence to show that adolescents' inherent characteristics are based on their brains' immaturity, the result of the interactions between the different cognitive functions – still in development – environmental demands, and the context in which these present themselves. Adolescent brains do not mature homogeneously and linearly, but instead develop according to cognitive and psychosocial abilities. This explains why adolescents might show developed abilities in certain contexts or scenarios, but not in others.

In scenarios where tasks are mainly cognitive – where there is time to make decisions, it is possible to consult an adult or evaluate the different choices and alternatives, and the level of stress is low – adolescents show competence levels similar to those of an adult (cognitive maturity) (Steinberg, 2009). More complex contexts – with high stress and emotional activation, pressure from peers, or little time to think – require coordination between affectivity and cognition (psychosocial maturity), which is still immature at the age of 16.

This temporal gap between the maturity of different abilities has generated legal debates, but it also establishes the different progressive capacities of adolescents under the law. In this sense, these different capacities establish the grounds as to why adolescents can make sanitary decisions and vote at 16 years, but cannot buy alcohol or cigarettes until later.

As Steinberg (2009) has pointed out, the cognitive maturity required for decision-making needs logical reasoning and the capacity for the comprehension and processing of relevant information. Following this line of thought, it can be seen that maturity in certain aspects of judgment develops between the ages of 11 and 16 years, arising from an improvement in abstraction, deliberation, and methods of induction. These cognitive abilities, which mature between the end of childhood and the middle of adolescence, reach a peak at the age of 16 years. That is to say, in "cold thinking" contexts, there is no significant difference in the capacity to comprehend and reason information in order to make decisions between middle adolescence and adulthood. As has been mentioned, this could lay the groundwork for the argument that the age of competence to make medical decisions should be 16 years.

However, it must be highlighted that only certain aspects of judgment mature around the age of 16 years, whereas some other cognitive–intellectual aspects are influenced by the affective sphere. In that sense, connections between the brain regions that integrate cognition and emotion are still immature during middle adolescence (16 years of age). This explains why adolescents show a less developed ability to exercise effective judgment in contexts where they find themselves influenced by emotional and social variables, despite their cognitive capacities.

Most antisocial behaviors in adolescents appear within the peer group (Piquero et al., 2003). They are mostly impulsive behaviors and are not premeditated. When adolescents are with a group of peers, unsupervised, and emotionally activated, they are more sensitive to focusing on short-term rewards and less able to think about negative consequences, which debilitates their competence to make reasonable decisions (Steinberg et al., 2009a). The influence of peer pressure is therefore more intense during adolescence than during adulthood (Gardner and Steinberg, 2005).

These characteristics, which are common signs of adolescents' immaturity, must be (and are) taken into account for the construction of public policies; there are, for example, special regulations that stipulate the age under which the sale of cigarettes and alcohol is prohibited, the minimum age for driving, and the age at which contraceptives can be accessed (Steinberg et al., 2009a). These policies can be improved in line with new scientific evidence. It has recently been recommended that the minimum age required for smoking should be raised (Farber et al., 2015), while other measures to prohibit adolescent drivers younger than 18 years from carrying passengers, or to limit their ability to do so according to the time of day, have also been suggested[4].

When the context allows time for adolescents to decide, consult, or obtain objective information about the risks, benefits, and alternative options, or when the influence of emotions and peers can be minimized, adolescents older than 16 years are bound to be able to make more deliberate and reasonable decisions in a similar capacity to adults (Steinberg et al., 2009a). Making decisions about health, giving medical consent to take part in an investigative project, voting, and making decisions with juridical consequences are examples of such scenarios.

Taking into account the diminished responses that adolescents have to aversive stimuli, public policies of containment should be developed to act over adolescents who experiment with risk in negative situations, given that it is less probable that they would attribute any negative results to the way that they behave (Reyna and Farley, 2006).

Differences between adolescence and adulthood are also rooted in the maturation process, and in brain, cognitive, and psychological development, while also presenting ground for new arguments that discuss a differentiated criminal treatment with less culpability for adolescents, and which take their immaturity into account.

We understand that there are some aspects that it has not been possible to explore to their fullest in this medium. One such aspect concerns the cognitive abilities required to be subjected to a full criminal trial, and how these change across different ages (Kivisto et al., 2011). From a legal point of view, and based on the progressive capacities of adolescents, from the age of 16 years onward, adolescents can make decisions about their health in the

[4] A study published in 2000 recommended restrictions on vehicle passengers if the driver is younger than 18 years. This was based on the fact that for 16- and 18-years-old drivers, the risk of fatal accidents increases when they drive with other passengers in the vehicle (Chen et al., 2000).

same way as an adult. Studies about adolescents' capacity to be put on trial show that a large proportion of those who are younger than 16 years experience difficulties with specific tasks of legal reasoning (Ficke et al., 2006) and in completely comprehending their rights and how to apply them. Likewise, their capacities are influenced by stress, suggestibility, and their intellectual level (Kassin, 2008). There is strong evidence that supports the idea that youngsters who are 12 years or younger have a less developed ability to comprehend and reason juridical information when compared to adolescents who are older than 16 years or adults with no psychological alteration (Ficke et al., 2006). To this effect, research has shown that 20% of adolescents between the ages of 14 and 15 years show deficient capacities comparable to adults who have no capacity to face trial for mental health reasons (Grisso et al., 2003).

When analyzing the development and maturation of adolescents, it is also important to consider the interaction between the biological and environmental aspects; examples include the impact of different factors such as poverty, stress, and traumatic situations (Auyero and Berti, 2013). Socioeconomic status is a relevant environmental factor that affects the functioning of the adolescent brain. In a recent systematic review of studies conducted with individuals between the ages of 13 and 25 years, Buckley et al. (2019) have presented evidence that socioeconomic status influences neural activation related to the processing of emotional and social stimuli. For example, negative experiences lead to a greater degree of responses, observable through the activation of the frontal cortex, in individuals with a lower socioeconomic status. Simultaneously, this review clarified that individuals with different socioeconomic statuses can show different behavioral responses even though their corresponding patterns of neural activation are similar. In any case, the manner in which socioeconomic status affects the functioning of the adolescent brain can be influenced by other factors. In this regard, a previous study has shown "that positive maternal parenting might ameliorate the negative effects of socioeconomic disadvantage on frontal lobe development (with implications for functioning) during adolescence" (Whittle et al., 2017).

In conclusion, we argue that research on the development of the adolescent brain does not provide definitive answers about the exact age required for different juridical purposes. Nonetheless, the current state of knowledge does allow for reflection on the development and maturation of adolescents and the implications for considering them criminally responsible. It also validates demands for a system that provides adolescents with greater protection and that favors their healthy integral development. In any case, although a specific minimum age is not evident, this study is disposed not to recommend lowering the age of criminal responsibility, but rather the opposite.

The relevance of building bridges of effective communication between scientific studies of human behavior, the law, and justice systems must be emphasized; this particular case concerns the relation between neuroscience and the justice system for adolescents. It is not possible to continue along parallel pathways when the issues that demand solutions are convergent. We also consider it necessary for neuroscientific analysis to be taken into consideration by jurists, and for relevant breakthroughs in other disciplines to be included in their studies. As time passes, it is important – even essential – to increase the multidisciplinary collaborations that lead to legislative approaches based on evidence and public policies with measurable indicators (e.g., through the use of neuroimaging). In other words, an ongoing connection between neuroscientific advances and the answers to social problems that have previously been addressed through the application of the law is urgently needed.

AUTHOR CONTRIBUTIONS

This article is an updated, extended version of a manuscript originally published in Spanish in the journal *Boletín Mexicano de Derecho Comparado* and authored by EM, EG-L, and LM-Q (see Mercurio et al., 2018; permission was granted by the journal). NL, JÁM, and JMM contributed new perspectives and material. All the listed authors amended the article and approved the final version.

REFERENCES

Albert, D., Chein, J., and Steinberg, L. (2013). The teenage brain: Peer influences on adolescent decision making. *Curr. Direct. Psychol. Sci.* 22, 114–120. doi: 10.1177/0963721412471347

Albert, D., and Steinberg, L. (2011). Judgment and decision making in adolescence. *J. Res. Adolesc.* 21, 211–224. doi: 10.1111/j.1532-7795.2010.00724.x

Altimus, C. M. (2017). Neuroscience has the power to change the criminal justice system. *eNeuro* 3:ENEURO.0362-16.2016. doi: 10.1523/ENEURO.0362-16.2016

American Psychological Association (1990). *Amicus Curiae Brief Filed in the US Court of Appeals for the Eighth Circuit in Hodgson v. Minnesota, 497 US 417*, Vol. 11. 1987, Washington, DC: American Psychological Association.

American Psychological Association (2005). *Amicus curiae Brief Filed in US Supreme Court in Roper v. Simmons, 543 US 551*.Washington, DC: American Psychological Association

Auyero, J., and Berti, M. F. (2013). *La violencia en Los Márgenes. Una maestra y un Sociólogo en el Conurbano Bonaerense [The Violence in the Margins. A Teacher and a Sociologist in the Buenos Aires Conurbation]*. Buenos Aires: Katz Editores.

Barbalat, G., Domenech, P., Vernet, M., and Fourneret, P. (2009). Approche neuroéconomique de la prise de risque à l'adolescence [Risk-Taking in Adolescence: A Neuroeconomics Approach]. *L'Encéphale* 1674:91. doi: 10.1016/j.encep.2009.06.004

Blakemore, S. J., and Choudhury, S. (2006). Development of the adolescent brain: implications for executive function and social cognition. *J. Child Psychol. Psychiatry* 47, 296–312. doi: 10.1111/j.1469-7610.2006.01611.x

Botdorf, M., Rosenbaum, G. M., Patrianakos, J., Steinberg, L., and Chein, J. M. (2017). Adolescent risk-taking is predicted by individual differences in cognitive control over emotional, but not non-emotional, response conflict. *Cogn. Emot.* 31, 972–979. doi: 10.1080/02699931.2016.1168285

Buckley, L., Broadley, M., and Cascio, C. N. (2019). Socio-economic status and the developing brain in adolescence: a systematic review. *Child Neuropsychol.* 25, 859–884. doi: 10.1080/09297049.2018.1549209

Burnett, S., Bault, N., Coricelli, G., and Blakemore, S. J. (2010). Adolescents' heightened risk-seeking in a probabilistic gambling task. *Cogn. Dev.* 25, 183–196. doi: 10.1016/j.cogdev.2009.11.003

Casey, B. J., Getz, and Galvan, A. (2008). The adolescent brain. *Dev. Rev.* 28, 62–77. doi: 10.1016/j.dr.2007.08.003

Casey, B. J., and Jones, R. M. (2010). Neurobiology of the adolescent brain and behavior. *J. Am. Acad. Child Adolesc. Psychiatry* 49, 1189–1285. doi: 10.1016/j.jaac.2010.08.017

Catley, P., and Claydon, L. (2015). The use of neuroscientific evidence in the courtroom by those accused of criminal offenses in England and Wales. *J. Law Biosci.* 2, 510–549. doi: 10.1093/jlb/lsv025

Chein, J. M., Albert, D., O'Brien, L., Uckert, K., and Steinberg, L. (2011). Peers increase adolescent risk taking by enhancing activity in the brain's reward circuitry. *Dev. Sci.* 14, F1–F10. doi: 10.1111/j.1467-7687.2010.01035.x

Chen, L.-H., Baker, S. P., Braver, E. R., and Li, G. (2000). Carrying passengers as a risk factor for crashes fatal to 16- and 17-Year-Old drivers. *JAMA* 283, 1578–1582. doi: 10.1001/jama.283.12.1578

Churchland, P. S. (2012). *Braintrust: What Neuroscience Tells Us About Morality*. Princeton: Princeton University Press.

Ciranka, S. K., and van den Bos, W. (2019). Social influence in adolescent decision making: a formal framework. *Front. Psychol.* 10:1915.

Cohen, A. O., and Casey, B. J. (2014). Rewiring juvenile justice: the intersection of developmental neuroscience and legal policy. *Trends Cogn. Sci.* 18, 63–65. doi: 10.1016/j.tics.2013.11.002

Currie, C., Roberts, C., Morgan, A., Smith, R., Settertobulte, W., Samdal, O., et al. eds (2004). *Young People's Health in Context: Health Behaviour in School-aged Children (HBSC) Study: International Report from the 2001/2002 Survey*, 1st Edn. Available online at: https://apps.who.int/iris/bitstream/handle/10665/107560/e82923.pdf (accessed March 2, 2020).

Dahl, R. E. (2004). Adolescent brain development: a period of vulnerabilities and opportunities. keynote address. *Ann. N. Y. Acad. Sci.* 1021, 1–22. doi: 10.1196/annals.1308.001

de Kogel, C. H., and Westgeest, E. J. M. C. (2015). Neuroscientific and behavioral genetic information in criminal cases in the Netherlands. *J. Law Biosci.* 2, 580–605. doi: 10.1093/jlb/lsv024

Dishion, T. J., and Tipsord, J. M. (2011). Peer contagion in child and adolescent social and emotional development. *Annu. Rev. Psychol.* 62, 189–214. doi: 10.1146/annurev.psych.093008.100412

Eaton, D. K., Kann, L., Kinchen, S., Shanklin, S., Ross, J., Hawkins, J., et al. (2008). Youth risk behavior surveillance—United States, 2007. morbidity and mortality weekly report. *Surveill. Summar.* 57, 1–131.

Escobar, M. J., Huepe, D., Decety, J., Sedeño, L., Messow, M. K., Baez, S., et al. (2014). Brain signatures of moral sensitivity in adolescents with early social deprivation. *Sci. Rep.* 4:5354. doi: 10.1038/srep05354

Farahany, N. A. (2015). Neuroscience and behavioral genetics in US criminal law: An empirical analysis. *J. Law Biosci.* 2, 485–509. doi: 10.1093/jlb/lsv059

Farber, H. J., Nelson, K. E., Groner, J. A., and Walley, S. C. (2015). Public policy to protect children from tobacco, nicotine smoke. *Pedriatrics* 136, 998–1007. doi: 10.1542/peds.2015-3109

Federal Bureau of Investigation (2003). *(US Dept. of Justice) Age-specific Arrest Rates, and Race-specific Arrest Rates. (for)Selected Offenses, 1993-2001 (p. 74)*. Washington, DC: Federal Bureau of Investigation.

Feldstein Ewing, E., Bjork, J. M., and Luciana, M. (2018). Implications of the ABCD study for developmental neuroscience. *Dev. Cogn. Neurosci.* 32, 161–164. doi: 10.1016/j.dcn.2018.05.003

Ficke, S. L., Hart, K. J., and Deardorff, P. A. (2006). The performance of incarcerated juveniles on the macarthur competence assessment tool-criminal adjudication (MacCAT-CA). *J. Am. Acad. Psychiatry Law Online* 34:360.

Galvan, A. (2010). Adolescent development of the reward system. *Front. Hum. Neurosci.* 4:2010. doi: 10.3389/neuro.09.006.2010

Galvan, A., Hare, T. A., Parra, C. E., Penn, J., Voss, H., Glover, G., et al. (2006). Earlier development of the accumbens relative to orbitofrontal cortex might underlie risk-taking behavior in adolescents. *J. Neurosci.* 26:6885. doi: 10.1523/JNEUROSCI.1062-06.2006

García-López, E. (2004). Edad penal y Psicología Jurídica. La necesidad de una respuesta social al adolescente infractor [Criminal Age and Legal Psychology. The need for a social answer to juvenile delinquency]. *Psicol. Am. Lat.* 2.

García-López, E., and Mercurio, E. N. (2019). *Psicopatología Forense y Justicia Restaurativa. Perspectivas desde el Neuroderecho. [Forensic Psychopathology and Restorative Justice. A view from Neurolaw]*. Ciudad de México: Instituto Nacional de Ciencias Penales.

Gardner, M., and Steinberg, L. (2005). Peer influence on risk taking, risk preference, and risky decision making in adolescence and adulthood: an experimental study. *Dev. Psychol.* 41, 625–635. doi: 10.1037/0012-1649.41.4.625

Giedd, J. N. (2004). Structural magnetic resonance imaging of the adolescent brain. *Ann. N. Y. Acad. Sci.* 1021, 77–85. doi: 10.1196/annals.1308.009

Giedd, J. N. (2008). The teen brain: insights from neuroimaging. *J. Adolesc. Health* 42, 335–343. doi: 10.1016/j.jadohealth.2008.01.007

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., et al. (1999). Brain development during childhood and adolescence: a longitudinal MRI study. *Nat. Neurosci.* 2, 861–863. doi: 10.1038/13158

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., et al. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proc. Natl. Acad. Sci. U.S.A.* 101, 8174. doi: 10.1073/pnas.0402680101

Graham v. Florida (2010). 560 U.S. 48, 130 S. Ct. 2011, 176 L. Ed. 2d 825 2010 U.S. LEXIS 3881. Available online at: https://www.supremecourt.gov/opinions/09pdf/08-7412modified.pdf

Grisso, T., Steinberg, L., Woolard, J., Cauffman, E., Scott, E., Graham, S., et al. (2003). Juveniles' competence to stand trial: a comparison of adolescents' and adults' capacities as trial defendants. *Law Hum. Behav.* 27, 333–363. doi: 10.1023/A:1024065015717

Guyer, A. E., Silk, J. S., and Nelson, E. E. (2016). The neurobiology of the emotional adolescent: From the inside out. *Neurosci. Biobehav. Rev.* 70, 74–85. doi: 10.1016/j.neubiorev.2016.07.037

Hartley, C. A., and Somerville, L. H. (2015). The neuroscience of adolescent decision-making. *Curr. Opin. Behav. Sci.* 5, 108–115. doi: 10.1016/j.cobeha.2015.09.004

Herrera, M., Caramelo, G., and Picasso, S. eds (2015). *Código Civil y Comercial de la Nación Comentado [Civil and Commercial Code of the Nation Commented]*, 1st Edn. Buenos Aires: Infojus.

Hodgson v. Minnesota (1990). *497 U. S.*

ICAP (2012). *Drinking Age in the World*. Karachi: International Center for Alcohol Policies.

Icenogle, G., Steinberg, L., Duell, N., Chein, J., Chang, L., Chaudhary, N., et al. (2019). Adolescents' cognitive capacity reaches adult levels prior to their psychosocial maturity: Evidence for a "maturity gap" in a multinational, cross-sectional sample. *Law Hum. Behav.* 43, 69–85. doi: 10.1037/lhb0000315

Jackson v. Hobbs (2012). *132 S. Ct. 548, 181 L. Ed. 2d 395, 565 U.S. 1013 2012*. Available online at: https://www.supremecourt.gov/qp/10-09647qp.pdf

Jones, O. D., Bonnie, R. J., Casey, B. J., Davis, A., Faigman, D. L., Hoffman, M., et al. (2014). *Law and neuroscience: Recommendations submitted to the President's Bioethics Commission. Faculty Scholarship at Penn Law, Paper 1439*. Pennyslvania.

Kambam, P., and Thompson, C. (2009). The development of decision-making capacities in children and adolescents: Psychological and neurological perspectives and their implications for juvenile defendants. *Behav. Sci. Law* 27, 173–190. doi: 10.1002/bsl.859

Kassin, S. M. (2008). The psychology of confessions. *Ann. Rev. Law Soc. Sci.* 4, 193–217.

Kelley, A. E., Schochet, T., and Landry, C. F. (2004). Risk Taking and Novelty Seeking in Adolescence: Introduction to Part I. *Ann. N. Y. Acad. Sci.* 1021, 27–32. doi: 10.1196/annals.1308.003

Kivisto, A. J., Moore, T. M., Fite, P. A., and Seidner, B. G. (2011). Future orientation and competence to stand trial: the fragility of competence. *J. Am. Acad. Psychiatry Law Online* 39:316.

Loeber, R., Farrington, D., and Redondo, S. (2011). La transición desde la delincuencia juvenil a la delincuencia adulta. *Rev. Española Invest. Criminol.* 9, 1–41.

Logue, S., Chein, J., Gould, T., Holliday, E., and Steinberg, L. (2014). Adolescent mice, unlike adults, consume more alcohol in the presence of peers than alone. *Dev. Sci.* 17, 79–85. doi: 10.1111/desc.12101

Luna, B., Paulsen, D. J., Padmanabhan, A., and Geier, C. (2013). The teenage brain: cognitive control and motivation. *Curr. Direct. Psychol. Sci.* 22, 94–100. doi: 10.1177/0963721413478416

Luna, B., and Wright, C. (2016). "Adolescent brain development: implications for the juvenile criminal justice system," in *APA Handbooks in Psychology. APA Handbook of Psychology and Juvenile Justice*, eds K. Heilbrun, D. DeMatteo,

and N. E. S. Goldstein (American Psychological Association), 91–116. doi: 10.1037/14643-005

McKewen, M., Skippen, P., Cooper, P. S., Wong, A. S. W., Michie, P. T., Lenroot, R., et al. (2019). Does cognitive control ability mediate the relationship between reward-related mechanisms, impulsivity, and maladaptive outcomes in adolescence and young adulthood? *Cogn. Affect. Behav. Neurosci.* 19, 653–676. doi: 10.3758/s13415-019-00722-2

Mercurio, E. N. (2012). *Cerebro y Adolescencia: Implicancias Jurídico-Penales [Brain and Adolescence: Criminal Legal Implication]*, 1st Edn. Buenos Aires: Ad-Hoc.

Mercurio, E. N. (2014). *Influencia de los Avances en Neurociencia en las Decisiones Judiciales en el Derecho Penal Juvenil [Influence of advances in neurosciences in judicial decisions in juvenile criminal law]. In Informes en derecho: Vol. 14. Informes en Derecho. Estudios de Derecho Penal Juvenil V [Reports in law: Vol. 14. Reports in Law. Studies of Juvenile Criminal Law V]*. Santiago de Chile: Defensoría Penal Pública [Public Criminal Advocacy].

Mercurio, E. N., García-López, E., and Morales Quintero, L. A. (2018). Psicopatología forense y Neurociencias: Aportaciones al Sistema de Justicia para Adolescentes [Forensic Psychopathology and Neuroscience: Contributions to the Justice System for Adolescents]. *Bol. Mexicano Derecho Comparado* 1, 931–971.

Miller v. Alabama (2012). *132 S. Ct. 2455, 567 U.S. 460, 183 L. Ed. 2d 407 2012*. Available online at: https://www.supremecourt.gov/opinions/11pdf/10-9646g2i8.pdf

Ministry of Justice and Human Rights Argentina (2014). *Código Civil y Comercial de la Nación*. Buenos Aires: Ministry of Justice and Human Rights Argentina.

Organization of American States [OAS] (1969). *American Convention on Human Rights, "Pact of San Jose"*. Costa Rica: OAS.

Østby, Y., Tamnes, C. K., Fjell, A. M., Westlye, L. T., Due-Tønnessen, P., and Walhovd, K. B. (2009). Heterogeneity in subcortical brain development: a structural magnetic resonance imaging study of brain maturation from 8 to 30 Years. *J. Neurosci.* 29:11772. doi: 10.1523/JNEUROSCI.1242-09.2009

Palminteri, S., Kilford, E. J., Coricelli, G., and Blakemore, S. J. (2016). The computational development of reinforcement learning during adolescence. *PLoS Comput. Biol.* 12:e1004953. doi: 10.1371/journal.pcbi.1004953

Pascual Urzúa, J. R. (2014). "Evolución filogenética y desarrollo ontogenético de las funciones cognitivas [Phylogenetic evolution and ontogenetic development of cognitive functions]," in *Neurociencia cognitiva [Cognitive neuroscience]*, ed. D. Redolar Ripoll (Madrid: Editorial Médica Panamericana), 201–230.

Paulsen, D., Platt, M., Huettel, S., and Brannon, E. (2011). Decision-making under risk in children, adolescents, and young adults. *Front. Psychol.* 2:72. doi: 10.3389/fpsyg.2011.00072

Petanjek, Z., Judaš, M., Šimic, G., Rašin, M. R., Uylings, H. B. M., Rakic, P., et al. (2011). Extraordinary neoteny of synaptic spines in the human prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13281–13286. doi: 10.1073/pnas.1105108108

Pfeifer, J. H., Masten, C. L., Moore, W. E. III, Oswald, T. M., Mazziotta, J. C., Iacoboni, M., et al. (2011). Entering adolescence: resistance to peer influence, risky behavior, and neural changes in emotion reactivity. *Neuron* 69, 1029–1036. doi: 10.1016/j.neuron.2011.02.019

Piquero, A. R., Farrington, D. P., and Blumstein, A. (2003). The criminal career paradigm. *Crime Justice* 30, 359–506. doi: 10.1086/652234

Poon, K. (2018). Hot and cool executive functions in adolescence: development and contributions to important developmental outcomes. *Front. Psychol.* 8:2311. doi: 10.3389/fpsyg.2017.02311

Prinstein, M. J., Brechwald, W. A., and Cohen, G. L. (2011). Susceptibility to peer influence: Using a performance-based measure to identify adolescent males at heightened risk for deviant peer socialization. *Dev. Psychol.* 47:1167. doi: 10.1037/a0023274

Quinn, P. D., and Harden, K. P. (2013). Differential changes in impulsivity and sensation seeking and the escalation of substance use from adolescence to early adulthood. *Dev. Psychopathol.* 25, 223–239. doi: 10.1017/s0954579412000284

Reyna, V. F., and Farley, F. (2006). Risk and rationality in adolescent decision making: implications for theory, practice, and public policy. *Psychol. Sci. Public Interest* 7, 1–44. doi: 10.1111/j.1529-1006.2006.00026.x

Roper v. Simmons (2005). *543 U. S.* Available online at: https://www.supremecourt.gov/opinions/04pdf/03-633.pdf

Sedletzki, V. (2016). *Minimum Ages and the Realization of Adolescents' Rights. Revision of the Situation in Latin America and the Caribbean*. New York, NY: UNICEF, 53–56.

Shulman, E. P., Harden, K. P., Chein, J. M., and Steinberg, L. (2014). The development of impulse control and sensation-seeking in adolescence: independent or interdependent processes? *J. Res. Adolesc.* 26, 37–44. doi: 10.1111/jora.12181

Smith, A. R., Chein, J., and Steinberg, L. (2014). Peers increase adolescent risk taking even when the probabilities of negative outcomes are known. *Dev. Psychol.* 50, 1564–1568. doi: 10.1037/a0035696

Smith, A. R., Rosenbaum, G. M., Botdorf, M. A., Steinberg, L., and Chein, J. M. (2018). Peers influence adolescent reward processing, but not response inhibition. *Cogn. Affect. Behav. Neurosci.* 18, 284–295. doi: 10.3758/s13415-018-0569-5

Somerville, L. H., Hare, T., and Casey, B. J. (2011). Frontostriatal maturation predicts cognitive control failure to appetitive cues in adolescents. *J. Cogn. Neurosci.* 23, 2123–2134. doi: 10.1162/jocn.2010.21572

Sowell, E. R., Thompson, P. M., Leonard, C. M., Welcome, S. E., Kan, E., and Toga, A. W. (2004). Longitudinal mapping of cortical thickness and brain growth in normal children. *J. Neurosci.* 24, 8223–8231. doi: 10.1523/JNEUROSCI.1798-04.2004

Spear, L. P. (2000). The adolescent brain and age-related behavioral manifestations. *Neurosci. Biobehav. Rev.* 24, 417–463. doi: 10.1016/S0149-7634(00)00014-2

Spear, L. P. (2013). Adolescent neurodevelopment. *J. Adolesc. Health* 52(2 Suppl. 2), S7–S13. doi: 10.1016/j.jadohealth.2012.05.006

Steinberg, L. (2004). Risk taking in adolescence: what changes, and why? *Ann. N. Y. Acad. Sci.* 1021, 51–58. doi: 10.1196/annals.1308.005

Steinberg, L. (2008). A social neuroscience perspective on adolescent risk-taking. *Dev. Rev.* 28, 78–106. doi: 10.1016/j.dr.2007.08.002

Steinberg, L. (2009). Adolescent development and juvenile justice. *Ann. Rev. Clin. Psychol.* 5, 459–485. doi: 10.1146/annurev.clinpsy.032408.153603

Steinberg, L. (2013). The influence of neuroscience on US Supreme Court decisions about adolescents' criminal culpability. *Nat. Rev. Neurosci.* 14, 513–518. doi: 10.1038/nrn3509

Steinberg, L., Cauffman, E., Woolard, J., Graham, S., and Banich, M. (2009a). Are adolescents less mature than adults?: Minors' access to abortion, the juvenile death penalty, and the alleged APA "flip-flop.". *Am. Psychol.* 64, 583–594. doi: 10.1037/a0014763

Steinberg, L., Graham, S., O'Brien, L., Woolard, J., Cauffman, E., and Banich, M. (2009b). Age differences in future orientation and delay discounting. *Child Dev.* 80, 28–44. doi: 10.1111/j.1467-8624.200801244.x

Steinberg, L., and Monahan, K. C. (2007). Age differences in resistance to peer influence. *Dev. Psychol.* 43, 1531–1543. doi: 10.1037/0012-1649.43.6.1531

Tamnes, C. K., Østby, Y., Fjell, A. M., Westlye, L. T., Due-Tønnessen, P., and Walhovd, K. B. (2010). Brain maturation in adolescence and young adulthood: regional age-related changes in cortical thickness and white matter volume and microstructure. *Cereb. Cortex* 20, 534–548. doi: 10.1093/cercor/bhp118

Toga, A. W., Thompson, P. M., and Sowell, E. R. (2006). Mapping brain maturation. *Trends Neurosci.* 29, 148–159. doi: 10.1016/j.tins.2006.01.007

Trezza, V., Campolongo, P., and Vanderschuren, L. J. (2011). Evaluating the rewarding nature of social interactions in laboratory animals. *Dev. Cogn. Neurosci.* 1, 444–458. doi: 10.1016/j.dcn.2011.05.007

Trucco, E. M., Colder, C. R., and Wieczorek, W. F. (2011). Vulnerability to peer influence: a moderated mediation study of early adolescent alcohol use initiation. *Addict. Behav.* 36, 729–736. doi: 10.1016/j.addbeh.2011.02.008

United Nations Committe on the Rights of the Child [CRC] (2016). *General Comment No. 20 (2016) on the Implementation of the Rights of the Child During Adolescence, 6 December 2016, CRC/C/GC/20*. Available online at: https://www.refworld.org/docid/589dad3d4.html (accessed March 2, 2020)

UNICEF (2014). *Aportes Para la Cobertura Periodística Sobre la Rebaja de la edad de Imputabilidad [Contributions for Journalistic Coverage on the Reduction of the Minimum Age of Criminal Responsibility]*. New York, NY: UNICEF Montevideo.

UNICEF (2017). *Para Cada Adolescente Una Oportunidad. Posicionamiento sobre Adolescencia [An Opportunity for every Teenager. Positioning on Adolescence]*. Argentina: UNICEF.

United Nations General Assembly [UNGA] (1959). *Declaration of the Rights of the Child, resolution 1386 (XIV) of 10 December 1959*. New York, NY: UNGA.

United Nations General Assembly [UNGA] (1989). *Convention on the Rights of the Child, resolution 44/25 of 20 November 1989*. New York, NY: UNGA.

Whittle, S., Vijayakumar, N., Simmons, J. G., Dennison, M., Schwartz, O., Pantelis, C., et al. (2017). Role of positive parenting in the association between neighborhood social disadvantage and brain development across adolescence. *JAMA Psychiatry* 74, 824–832. doi: 10.1001/jamapsychiatry.2017.1558

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

OPINION
published: 25 February 2020
doi: 10.3389/fpsyg.2020.00302

Neuroscience in Youth Criminal Law: Reconsidering the Measure of Punishment in Latin America

Nicolás Ezequiel Llamas and José Ángel Marinaro**

Department of Law and Political Science, National University of La Matanza, San Justo, Argentina

Keywords: neurolaw, youth criminal law, juvenile crime and delinquency, juvenile criminal behavior, juvenile criminal law, capital punishment, life imprisonment, life imprisonment without parole

INTRODUCTION

Due to the new discoveries and advances made in technology in the field of neuroscience in the last few decades, it has been possible to get a better understanding of the development of the human brain. This has had a significant impact on youth criminal law, especially in relation to the behavior of adolescents and their capacity to control impulsive reactions.

In this article, we will discuss the repercussions of this improved understanding on the amount of penalty for convicted adolescents in Latin America.

It is important to mention that the minimum age of criminal responsibility on each country of this region is quite different (mostly between 12 and 16 years old). Despite this and other divergences, we think it is possible to make an approach from the point of view of the Inter-American Human Rights System.

MEASURE OF PUNISHMENT: COMPARATIVE DISPROPORTION

It could be argued that the majority of actions or omissions which constitute a crime in a certain country usually also constitute a crime in most countries around the world. However, the measure of the punishment that could be imposed as result of that same crime does not follow this generalization. In this respect, for example, there are several countries that do not impose capital punishment or life imprisonment.

In this context, and according to Comparative Law, we find large disparities between the penalties applied in different countries by the youth criminal law, the body of law that regulates crimes committed by a person under the age of majority. Latin American countries are a great example of this situation: while Brazil has a maximum penalty of 3 years of imprisonment for any crime committed by an adolescent between 12 and 18-year-old (Law 8069 [*Estatuto da Criança e do Adolescente*], s. 121), other countries, like Bahamas [*Penal Code*, s. 263 (3)] allow capital punishment. More examples are shown in **Figure 1**.

In order to analyze this correctly, we propose to classify the different legislative methods into three groups. First, there are legal systems that allow the transfer of young offenders to a criminal court (also known as "trial as an adult"). Second, there are those that allow the juvenile court to impose an adult sentence. Third, there are those that only allow juvenile sentences for young offenders, which are considerably less severe than adult sentences.

The first method is common in countries that have adopted the legal system known as Common Law (pure or mixed). The decision to transfer a young offender may contemplate several factors, but the most important ones are the severity of the offense and the age of the offender. This decision

OPEN ACCESS

Edited by:
Elena Rusconi,
University of Trento, Italy

Reviewed by:
Amedeo Santosuosso,
University of Pavia, Italy

****Correspondence:***
Nicolás Ezequiel Llamas
nicolasllamas@hotmail.com
José Ángel Marinaro
joseangelmarinaro@yahoo.com.ar

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *28 August 2019*
Accepted: *07 February 2020*
Published: *25 February 2020*

Citation:
Llamas NE and Marinaro JÁ (2020) Neuroscience in Youth Criminal Law: Reconsidering the Measure of Punishment in Latin America. Front. Psychol. 11:302. doi: 10.3389/fpsyg.2020.00302

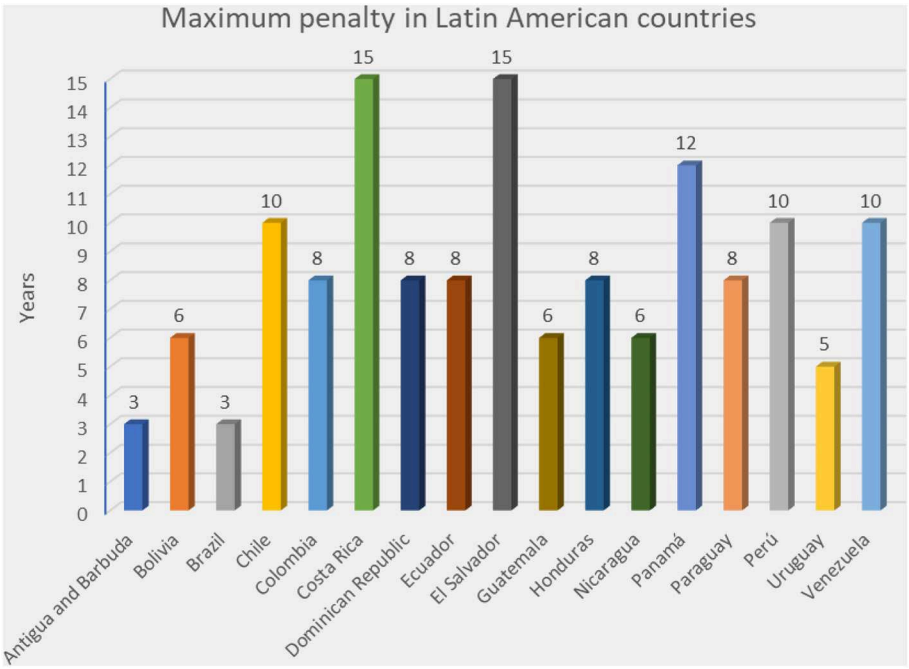


FIGURE 1 | This figure shows the maximum penalty in some Latin American countries that may be imposed on adolescents (as per references listed below). Countries that allow life imprisonment or capital punishment have been excluded for not been able to be shown. Antigua and Barbuda: *Child Justice Act* (No. 23 of 2005), c. X, s. 69(2); Bolivia: Law 548 [*Código Niña, Niño y Adolescente*], s. 268; Brazil: Law 8069 [*Estatuto da Criança e do Adolescente*], s. 121; Chile: Law 20084 [*Sistema de Responsabilidad de los Adolescentes por Infracciones a la Ley Penal*], s. 18; Colombia: Law 1098/2006 [*Código de la Infancia y la Adolescencia*], s. 187; Costa Rica: Law 7576 [*Ley de Justicia Penal Juvenil*], s. 131; Dominican Republic: Law 136-03 [*Código para la protección de los derechos de los Niños, Niñas y Adolescentes*], s. 340; Ecuador: Law 100 [*Código de la Niñez y Adolescencia*], s. 358(3); El Salvador: Law 869 [*Ley del Menor Infractor*], s. 15-17; Guatemala: Law 27/2003 [*Ley de Protección integral de la niñez y adolescencia*], s. 252(b); Honduras: Law 73/96 [*Código de la Niñez y la Adolescencia*], s. 205; Nicaragua: Law 287/1998 [*Código de la Niñez y la Adolescencia*], s. 206; Panamá: Law 40/1999 [*Régimen Especial de Responsabilidad Penal para la Adolescencia*], s. 141; Paraguay: Law 1680/2001 [*Código de la Niñez y la Adolescencia*], s. 207; Perú: Law 27337 [*Código de los Niños y Adolescentes*], s. 235; Uruguay: Law 17823 [*Código de la Niñez y la Adolescencia*], s. 91; Venezuela: Law 5859 [*Ley Orgánica para la Protección del Niño, Niña y Adolescente], s. 628*. Canada (*Youth Criminal Justice Act* [S.C. 2002, c. 1, s. 64(1)] and Grenada [*Juvenile Justice Act*, Act No. 24 of 2012, s. 4(2)] may impose life imprisonment. Through the reports called "Concluding observations" made by the United Nations Committee on the Rights of the Child (available on https://www.ohchr.org/SP/Countries/LacRegion/Pages/LacRegionIndex.aspx) we were able to establish that the death penalty could be imposed in Bahamas and Saint Lucia; and life imprisonment could be imposed in Barbados, Belize, Cuba, Dominica, Guyana, Haiti, Jamaica, Mexico, Saint Kitts and Nevis, Saint Vincent and the Grenadines, Suriname and Trinidad and Tobago.

may be made by a judge (judicial waiver), a prosecutor (prosecutorial discretion), or by the law itself (statutory exclusion).

The second method is mostly used in countries that have adopted the civil law system. Like the prior one, the severity of the offense and the age of the offender are the main factors used to make the decision. Despite their differences, both systems enable the sentencing of a young offender as an adult.

The third one, however, prohibits that kind of penalty, which also means it prohibits capital punishment and life imprisonment. It is not possible to make any other assumption about this matter since the maximum amount of penalty is quite diverse in every jurisdiction.

Additionally, there are international laws that prohibit capital punishment or life imprisonment for young offenders, like article 37 of the United Nations Convention on the Rights of the Child, and human-rights courts that do not allow for adolescents to be sentenced with the same punishment that may be imposed on an adult, like the leading case "Mendoza et al. v. Argentina, Preliminary Objections, Merits, and Reparations, Judgment" of the Inter-American Court of Human Rights (ser. C, No. 260, May 14, 2013).

OVERVIEW OF NEUROLAW REGARDING ADOLESCENTS

Having said all that, it should be affirmed that several neuroscientific studies have proved that adolescents do not have the same cognitive capacity as an adult. In particular, it has been suggested that the frontal lobe, whose functions involve controlling and judging impulse and risk, projecting future consequences resulting from current actions (Fuster, 2001; Martinez Selva et al., 2006), continues its development well into young adulthood (Gogtay et al., 2004; Giedd, 2008).

Thus, disadvantageous decision making and risky behavior shown by adolescents are considered to be related to the slower developing prefrontal cortex (Smith et al., 2012), which

has been linked to prominent differences in cognitive capacity (Cauffman and Steinberg, 2000; Galvan et al., 2006; Eshel et al., 2007). Further investigations have been made, some of them related to drug abuse or peer influence, which support this matter (Blakemore, 2012; Spear, 2013; Brizio et al., 2015; van Duijvenvoorde et al., 2016)[1].

The impact of those studies was meaningful for the judiciary system of the United States since they were used by its Supreme Court to sentence the leading cases *Roper v. Simmons* (543 U.S. 551), *Graham v. Florida* (560 U.S. 48), and *Miller v. Alabama* (567 U.S. 460). In addition, there is an ongoing debate about their legal implications (Steinberg, 2009; Delmage, 2013).

DISCUSSION

The Supreme Court of the United States stated that "a sentence lacking any legitimate penological justification is by its nature disproportionate to the offense" (*Graham v. Florida*, 560 U.S. 48, p. 20). However, it is necessary to analyze if penological justifications designed for adults are applicable to juveniles.

This implies a change in basic assumptions. Penological justifications have been created and built on suppositions tied up with notions of agency, freedom, and free will. Whenever a sentence requires a person acting purposely, the lack of intent means there is absence of blameworthiness as well as absence of any justification for condemning. Therefore, if it is proved that adolescents do not have the same capacity as an adult to observe the law, it does not only impact on the personal culpability but also the assumptions of the penological justification itself.

In this regard, equality and non-discrimination before the law should not only be considered as giving the same legal treatment to all human beings in general, but also to give different treatment to those who are not equals. Consequently, applying similar punishment to juvenile offenders and adult offenders for the same crime should be judged incompatible with legal principles and also with the current state of the science.

In modern criminal law there is no debate that any sentence must take into consideration the moral responsibility of the perpetrator. However, this same principle wrongly causes controversy when the outcome of its application consists of a reduction in culpability, and therefore in the size of the imposed penalty.

For all these reasons, we consider that any law or jurisprudence that makes the transfer of a juvenile offender to an adult court possible, or allows an adult sentence to be imposed on them, ought to be reconsidered. As it was mentioned before, there are many countries whose legislation provides considerable differences between juvenile and adults offenders, and the Inter-American Court of Human Rights has represented an important step in this direction (Llamas, 2019).

Nevertheless, the legal impact of the neuroscientific findings and technologies is an open debate (Muñoz Ortega, 2013, 2018). Nowadays, we are observing an exponential increase of publications about adolescents and their behavior related to alcohol, drugs, stress, and peer influence, among other topics. Some of them even suggest that the age of 18 is not a scientifically correct watershed between adolescent and adult criminal responsibility (Mercurio, 2012; Mercurio et al., 2019).

The topic is crucial when considering some countries with high levels of poverty and malnutrition in childhood, which may affect the development of the human brain and its cognitive abilities (Mercurio, 2016), as well as the known effects of deprivation (Llamas and Marinaro, 2017)[2].

As a final reflection, we want to mention that some very old Spanish laws, which were in force long before the independence of Latin-American countries (López de Guevara, 1843), did not allow adolescents to be sentenced as adults. In a way, it seems that new discoveries might prove scientifically what was presumed righteous long ago.

AUTHOR CONTRIBUTIONS

NL and JM wrote the manuscript with equal contributions.

FUNDING

This work was supported by the National University of La Matanza, Argentina.

ACKNOWLEDGMENTS

We thank José Manuel Muñoz Ortega for comments and useful discussions regarding earlier versions of this manuscript. We also thank Carolina Valeria De Valois for the translation work.

[1]There are other aspects of the development of prefrontal cortex which might play a major role in terms of behavioral outcomes, such as hormonal influences onto the brain (Blakemore et al., 2010).

[2]It is important to remark that there are cultural and ethical non-scientific aspects that were at the basis of the exclusion of the most severe punishments for young people in Europe and in the United States.

REFERENCES

Blakemore, S.-J. (2012). Imaging brain development: the adolescent brain. *Neuroimage* 61, 397–406. doi: 10.1016/j.neuroimage

Blakemore, S.-J., Burnett, S., and Dahl, R. E. (2010). The role of puberty in the developing adolescent brain. *Human Brain Mapp.* 31, 926–933. doi: 10.1002/hbm.21052

Brizio, A., Gabbatore, I., Tirassa, M., and Bosco, F. M. (2015). "No more a child, not yet an adult:" studying social cognition in adolescence. *Front. Psychol.* 6:1011. doi: 10.3389/fpsyg.2015.01011

Cauffman, E., and Steinberg, L. (2000). (Im) maturity of judgment in adolescence: why adolescents may be less culpable than adults. *Behav. Sci. Law* 18, 741–760. doi: 10.1002/bsl.416

Delmage, E. (2013). The minimum aof criminal responsibility: a medico-legal perspective. *Youth Justice* 13, 102–110. doi: 10.1177/1473225413492053

Eshel, N., Nelson, E. E., Blair, R. J., Pine, D. S., and Ernst, M. (2007). Neural substrates of choice selection in adults and adolescents: development of the ventrolateral prefrontal and anterior cingulate cortices. *Neuropsychologia* 45, 1270–1279. doi: 10.1016/j.neuropsychologia.2006.10.004

Fuster, J. M. (2001). The prefrontal cortex—an update: time is of the essence. *Neuron* 30, 319–333. doi: 10.1016/S0896-6273(01)00285-9

Galvan, A., Hare, T. A., Parra, C. E., Penn, J., Voss, H., Glover, G., et al. (2006). Earlier development of the accumbens relative to orbitofrontal cortex might underlie risk-taking behavior in adolescents. *J. Neurosci.* 26:6885. doi: 10.1523/JNEUROSCI.1062-06.2006

Giedd, J. N. (2008). The teen brain: insights from neuroimaging. *J. Adolesc. Health* 42, 335–343. doi: 10.1016/j.jadohealth.2008.01.007

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., et al. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proc. Natl. Acad. Sci. U.S.A.* 101, 8174–8179. doi: 10.1073/pnas.0402680101

Llamas, N. E. (2019). *La pena en el derecho penal juvenil [The penalty in youth criminal law] (Tesis doctoral)*. Universidad Nacional de La Matanza, Escuela de Posgrado, San Justo, Argentina.

Llamas, N. E., and Marinaro, J. Á. (2017). La hipótesis Cajal-Hebb: Vinculación entre las deprivaciones de los jóvenes delincuentes y las neurociencias [The Cajal-Hebb hypothesis: link between the deprivations of young offenders and neurosciences]. *Rev. Derecho Penal Criminología (La Ley)* VII, 203–209.

López de Guevara, G. (ed.). (1843). *Las Siete Partidas del Sabio Rey Don Alfonso IX [The Seven Divisions of Law of the Wise King Don Alfonso IX]*. Available online at: http://www.cervantesvirtual.com/obra/las-siete-partidas-del-sabio-rey-don-alfonso-el-ix-sic

Martinez Selva, J. M., Sanchez Navarro, J. P., Bechara, A., and Román Lapuente, F. (2006). Mecanismos cerebrales de la toma de decisiones [Brain mechanisms involved in decision-making]. *Rev. Neurol.* 42, 411–418. doi: 10.33588/rn.4207.2006161

Mercurio, E. N. (2012). *Cerebro y Adolescencia: Implicancias Jurídico-Penales [Brain and Adolescence: Legal-Criminal Implications], 1st Edn.* Buenos Aires: Ad-Hoc.

Mercurio, E. N. (2016). Pobreza y discapacidad intelectual en el sistema penal: los invisibles [Poverty and intellectual disability in the criminal system: the invisible]. *VERTEX Rev. Argentina Psiquiatría* XXVII, 197–207.

Mercurio, E. N., García López, E., and Morales Quintero, L. A. (2019). Psicopatología forense y neurociencias: aportaciones al sistema de justicia para adolescentes [Forensic phsychopathology and neuroscience: contributions to the justice system for adolescents]. *Bol. Mexicano Derecho Comparado* 1, 931–971.

Muñoz Ortega, J. M. (2013). Neurofilosofía y libre albedrío [Neurophilosophy and free will]. *Daimon Rev. Int. Filosofia* 59, 57–70.

Muñoz Ortega, J. M. (2018). Mental causation and neuroscience: the semantic pruning model. *Theoria Int. J. Theory Hist. Found. Sci.* 33, 379–399. doi: 10.1387/theoria.17312

Smith, D. G., Xiao, L., and Bechara, A. (2012). Decision making in children and adolescents: impaired iowa gambling task performance in early adolescence. *Dev. Psychol.* 48, 1180–1187. doi: 10.1037/a0026342

Spear, L. P. (2013). Adolescent neurodevelopment. *J. Adolesc. Health* 52, S7–S13. doi: 10.1016/j.jadohealth.2012.05.006

Steinberg, L. (2009). Should the science of adolescent brain development inform public policy? *Am. Psychol.* 64, 739–750. doi: 10.1037/0003-066X.64.8.739

van Duijvenvoorde, A. C. K., Peters, S., Braams, B. R., and Crone, E. A. (2016). What motivates adolescents? neural responses to rewards and their influence on adolescents' risk taking, learning, and cognitive control. *Adolesc. Brain* 70, 135–147. doi: 10.1016/j.neubiorev.2016.06.037

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 June 2019
doi: 10.3389/fpsyg.2019.01483

Conceptualizations of Addiction and Moral Responsibility

Jostein Rise and Torleif Halkjelsvik*

Department of Alcohol, Tobacco and Drugs, Norwegian Institute of Public Health, Oslo, Norway

The present study explored the connection between conceptualizations of addiction and lay people's inferences about moral responsibility. In Study 1, we investigated how natural variations in people's views of addiction were related to judgments of responsibility in a nationwide sample of Norwegian adults. In Study 2, respondents recruited from Mechanical Turk were asked to consider different conceptualizations of addiction and report on how these would affect their judgments of moral responsibility. In Study 3, we tested whether manipulating conceptualizations through textual information and through the framing of addiction in terms of states versus behavior could influence participants' judgments of moral responsibility. We found that attributions of moral responsibility were lower when addiction was connected to diseases and disorders, such as dysfunctional processes in the brain, and greater when addiction was associated with agency and addictive behaviors. In conclusion, different conceptualizations of addiction imply different moral judgments, and conceptualizations are malleable.

Keywords: addiction, agency, free will, moral judgment, conceptualizations, responsibility

OPEN ACCESS

Edited by:
José M. Muñoz,
Universidad Europea de Valencia, Spain

Reviewed by:
Andrew Vonasch,
University of Canterbury, New Zealand
Marta Miquel,
University of Jaume I, Spain

****Correspondence:***
Jostein Rise
jostein.rise@fhi.no;
josteinorise@gmail.com

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *28 March 2019*
Accepted: *11 June 2019*
Published: *28 June 2019*

Citation:
Rise J and Halkjelsvik T (2019) Conceptualizations of Addiction and Moral Responsibility. Front. Psychol. 10:1483. doi: 10.3389/fpsyg.2019.01483

Addiction as a phenomenon is a puzzle, paradox, and slippery concept for which definitions and classifications in diagnostic systems have changed with cultural, political, and scientific developments (Berridge et al., 2014; Room et al., 2015). A key concept in the discourse about addiction is the question about moral responsibility (see e.g., Morse, 2004; Foddy, 2011; Levy, 2011; Uusitalo, 2011). Are individuals addicted to substances morally responsible for their use, or does addiction represent a form of involuntary behavior? The starting point of the present study is that the answer to this question depends on the way addiction is conceptualized, that is, how people view and describe addictions. Knowledge of the connection between judgment of moral responsibility and conceptualizations of addiction may be important for understanding, and potentially changing, how addicted individuals are treated in society. Moral judgments may have a range of consequences from how drug policies are formed to how professionals in the healthcare and criminal justice systems behave toward addicted persons (cf. Pickard and Pierce, 2013). The concept of moral responsibility has been treated in diverse bodies of literatures, from which three lines of inquiry may be particularly relevant: scientific addiction models, research on stigma and attribution, and the contemporary literature on free will and agency.

ADDICTION MODELS

The scientific discourse about addiction has been dominated by two models: the disease model and the choice model (Morse, 2004; Henden et al., 2013; Uusitalo et al., 2013). The former considers addiction as following a disease-like course, with behaviors that have taken control of the person–so-called compulsive actions. A modern version of the disease model is the

view of addiction as a brain disease (see e.g., Kennett and McConnell, 2013). The brain disease model holds that neural processes and chemical reactions following repeated intake of drugs cause lasting brain changes so that the reward system is hijacked and governs the motivations behind addictive behaviors. This model has recently been challenged from a number of perspectives (see Heyman, 2009; Henden et al., 2013; Lewis, 2015; Heather et al., 2017; Pickard, 2017a). In contrast to the brain disease model, the choice model holds that addictive behaviors are governed by universal principles of choice and motivation. The choice model has been referred to as the successor of the moral model of addition (Kennett and McConnell, 2013), where addiction was considered a moral failure and addicts could be perceived as people of bad character (see Pickard, 2017b). However, moral considerations are not core features of modern choice theories of addiction (cf. Heyman, 2009).

Recently, several authors have argued in favor of views that place addiction somewhere in the middle of a continuum between nonvoluntary behavior and voluntary actions (Henden et al., 2013; Holton and Berridge, 2013; Heather, 2017a). This middle ground involves *excusing conditions* for addictive behaviors, meaning that there are strong forces at play that are difficult, but not impossible, to resist (see Morse, 2004; Levy, 2011; Pickard and Pierce, 2013).

STIGMA AND ATTRIBUTION

Individuals addicted to drugs are heavily stigmatized and viewed by lay people as more dangerous and blameworthy than individuals with mental illness or physical disabilities (Corrigan et al., 2009). Several factors appear to moderate the level of stigmatization (e.g., Corrigan et al., 2001, 2002; Pinfold et al., 2003; Schulze et al., 2003). For instance, in a recent survey, stigmatization of people with drug addiction was influenced by factors related to the stigmatized person (such as gender, age, and duration of addiction) and demographic characteristics of the person making the judgment (Sattler et al., 2017). The authors of the study found their results to be fairly consistent with Weiner's attribution theory (e.g., Weiner, 1995, 2006). A core assumption in the attribution theory is that controllability of a stigmatized behavior is consequential for perceived responsibility, which, in turn, is consequential for social emotions and outcomes such as helping behavior. Thus, perception of responsibility plays a central role in a process linking inferences regarding causes and controllability to emotional and behavioral consequences (Weiner, 1995; see also Shaver, 1985).

FREE WILL

Lay people seem to associate addiction with a loss of free will (Vonasch et al., 2017). Because free will is held to be a prerequisite for an agent to be punished for wrongdoing and praised for doing well, a number of scholars have posited a close relation between free will and moral responsibility (see Nahmias, 2018). The main debate in philosophy revolves around whether free will and moral responsibility are compatible with determinism–the idea that whatever happens is fully determined (caused) by previous events and the laws of nature (Mele, 2006). While *compatibilists* hold that free will and moral responsibility are compatible with determinism, *incompatibilists* deem that if determinism is true, then humans cannot have free will and be morally responsible for their actions. Results regarding this issue from empirical research on lay intuitions are divided (Cova and Kitano, 2014). While Nahmias and Murray (2010) claim that ordinary people are natural compatibilists, Nichols and Knobe (2007) claim that they are natural incompatibilists.

A recent psychological model of free will does not focus on whether or not lay people believe in free will but on what they mean by free will (Monroe and Malle, 2010, 2015). In essence, free will means that choices are unconstrained by internal and external circumstances (Monroe and Malle, 2010, 2015; Feldman et al., 2014; Vonasch et al., 2018). In one study, Monroe et al. (2016) found that after accounting for perceived choice capacities, nothing was left for a general and abstract belief in free will to account for lay peoples' judgment of an agent's immoral behavior. This suggested that a general belief in free will is a shorthand lay people use for the ascription of these capacities.

Thus, lay people's ascription of moral responsibility associated with addiction can be placed on a continuum from low to high (cf. Sinnott-Armstrong, 2013), and underlying this continuum is a model of freedom of action and free will as capacities to make decisions and exercise control. By this account, one should not treat free will and freedom of action as all-or-nothing properties (Nahmias, 2018). Of particular interest for the present research are the results from experimental philosophy studies on free will and determinism demonstrating that lay people's responses to questions of moral responsibility can vary dramatically depending on the way researchers formulate the scenario (see Cova and Kitano, 2014 for a review).

CONCEPTUALIZATIONS OF ADDICTION AND MORAL RESPONSIBILITY

As part of the 2012 Queensland Social Survey in Australia, Meurk et al. (2014) found that considering addiction as a brain disease or as an ordinary disease did not affect beliefs about stigma nor belief about the use of coerced treatment on and imprisonment of heroin users. Furthermore, the respondents' views on causes of addiction were inconsistent predictors of these beliefs. Meurk et al. (2014) argued that these results corroborated those of their prior qualitative studies, indicating that new information about addiction, in particular information portraying addiction as a brain disease, would not produce dramatic shifts in people's beliefs about addiction. Similarly, Rather (1991) investigated lay models of alcohol addiction and reported no effect of the manipulation of a disease model versus social-learning model of alcohol addiction on attitudes toward alcoholics or judgments of

deservingness of help, even though the manipulation affected beliefs about the causes of addiction.

The above studies did not directly concern moral judgments but hinted at the difficulties in linking conceptualizations of addiction to attributions of moral responsibility among lay people, at least in terms of changing such conceptualizations. In a recent and highly relevant study, Racine et al. (2017) compared the effect of three types of neuroscientific information about addiction (alcohol and cocaine) on the attribution of free will: (1) a textual neuroscience description of addiction, (2) neuroimages of a nonaddict's and of an addict's brain, (3) a combination of text and neuroimages, and (4) a control condition with no information. A factor analysis of a scale measuring lay beliefs about whether addicts have free will revealed two distinct free will factors denoted Responsibility and Volition. One hypothesis was that a neuroscience perspective of addiction would reduce the attribution of free will and subsequently the blame. However, they only found a significant effect of the combined image and textual description on the volition subscale in terms of diminished free will for cocaine addiction. Racine et al. (2017) argued that the results indicate that naturally occurring neuroscientific information about addiction might have limited effects on attributions of free will (responsibility and volition), and, accordingly, that the merits of the brain disease model may have been overstated.

The above studies involved efforts to change conceptualizations of addiction. Another question in the literature of addiction is whether, and how, natural variations in the lay peoples' conceptualizations matter for moral judgments. Research on perceptions of addictions to different types of behavior suggests that the type of addiction (i.e., type of substance) is consequential for moral judgments. In a nationwide study among Swedish adults, Blomqvist (2009) explored responsibility judgments for nine different types of addiction. He distinguished between responsibility for the onset of a problem and responsibility for solving the problem (see Brickman et al., 1982). Addiction to tobacco fit into a *moral model*, where lay people perceive users as responsible for both the onset and the solution to the problem. Addiction to alcohol, sedatives, and cannabis were placed within the *compensatory model*, where users are responsible for the solution of the problem but not the onset. Hard drug addicts fit into a combination of the *medical model* (neither responsible for the onset nor for its solution, i.e., they have a disease and should receive treatment) and the *enlightenment model* (responsible for the problem but not for the solution of the problem), implying that addicts are victims that need external help to overcome the addiction. The study by Blomqvist (2009) and similar studies (Halkjelsvik and Rise, 2014; Rise et al., 2014) suggest that conceptualizations of addiction, in particular those connected to beliefs about the causes of addiction (see also Weiner et al., 1988), can be consequential for moral judgments.

THE PRESENT RESEARCH

The issue of how conceptualizations of addiction are linked to moral responsibility can be approached in several ways. When a person holds a certain view of addiction, what does this entail in terms of moral judgments? When a person receives and accepts a certain description of addiction, what does he/she believe this implies in terms of moral responsibility? Can information or the framing of addiction shape people's own beliefs regarding moral responsibility? These are different questions, but they all pertain to the relation between conceptualizations of addiction and moral judgments. When one describes addiction in research and in the media, knowledge about what the different labels and descriptions imply in terms of moral judgments can be valuable, particularly if the words have an impact on other people's moral judgments.

We explored the connections between conceptualizations of addiction and moral judgments in three studies, using different approaches. In Study 1, we recruited a broad sample of the Norwegian population and used a wide array of textual descriptions reflecting the ways addiction has been described in the literature. We investigated how variations in people's endorsement of these descriptions related to their judgments of responsibility. In Study 2, we explicitly asked people to accept different conceptualizations and then investigated how this would affect judgments of moral responsibility. In Study 3, we tested whether we were able to manipulate judgments of responsibility through textual information about addiction and through changing the object of evaluation by framing addiction in terms of addictive states versus addictive behaviors.

Study 1

People have different backgrounds, values, and ideologies, and it is reasonable to assume that there is substantial variation in people's views of addiction. The same individual can have different views of addiction, depending on the type of addictive behavior involved (e.g., cigarette smoking versus use of heroin). In Study 1, we attempted to exploit this natural variation in conceptualizations of addictions by exploring lay people's ratings of a range of addiction descriptors that were derived from the scientific literature on addiction.

Each respondent rated several types of addiction which enabled us to explore two types of effects in Study 1, one based on between-person differences in conceptualizations of addiction and another based on within-person differences. Both effects may be informative regarding the relation between conceptualizations and moral judgments; however, only the latter removes time-invariant confounding. We controlled for the overall differences between types of addiction (i.e., the averages across the sample of individuals), as these may be heavily influenced by the legal status and the prevalence of the addictive behavior.

Methods

Data and Sample

The recruitment panel of an independent research company was used to invite a representative sample of Norwegians aged 20–70 with access to the internet (i.e., the online population). Of the 2,964 invited to participate, 2,037 responded to at least

one question in a large survey on addictions and related issues. Except for one analysis (N = 1,853), the number of respondents ranged from 1979 to 2011 in the statistical analyses. The mean age of the sample was 47, SD = 14; 50% were women. Results from the same survey have previously been reported in Melberg et al. (2013), Rise et al. (2014, 2015), but none of the present analyses have been published before. None of the studies reported in the present article required ethics approval per our institution's guidelines and Norwegian law. We did not collect IP addresses or any personal or sensitive information. Participation was voluntary; participants were informed that their responses would be used in research; and they were asked to consent by proceeding to the survey questions.

Measures

Types of Addiction

The study involved ratings of addiction to cocaine, hashish (cannabis), alcohol, gambling, smoking, amphetamine, sedatives, snus (Swedish moist snuff)], and heroin. Participants rated all nine addiction types in terms of 13 different addiction descriptors.

Addiction Descriptors

After the initial text: "Addiction to [type of addiction] is/represents…", respondents rated their level of agreement with 13 different descriptors of addiction (see **Table 1**) on a seven-point scale from "Fully disagree" (coded 1) to "Fully agree" (coded 7). As an example, the participants rated the level of agreement with the statement "Addiction to Cocaine is/represents…reduced willpower". The descriptors were based on an informal survey of the literature and reflected conceptualizations by lay people and scientists (see Rise et al., 2015).

Responsibility Judgments

The outcome measure was ratings of whether a person addicted to (type of addiction) should be held responsible for becoming addicted to the substance/behavior. The response scale was from "To a very small extent" (coded 1) to "To a very large extent" (coded 7). The option "do not know" was coded as missing.

Statistical Analyses

Analyses were performed in STATA 14.1 using the "mixed" command with maximum likelihood estimation and robust standard errors; p's were based on the default large-sample tests. We ran separate regression models for each of the 13 addiction descriptors in **Table 1**. In each analysis, the outcome measure was a variable comprising the responsibility ratings for all the nine types of addiction. Type of addiction was controlled for by dummy indicators, and the predictor of interest was the endorsement of the given descriptor (i.e., the extent to which respondents agreed that a descriptor is/represents a given addiction). We included two different terms in the regressions to estimate the effect of endorsement of a given descriptor on moral judgments. One term represented the between-individuals effect and was estimated by a variable consisting of each individual's mean endorsement of the given descriptor across addiction types; another term represented the within-individual effect and was estimated by the endorsement ratings minus the respective individual's mean endorsement (for more on this "within-between" approach, see Bell and Jones, 2015; Snijders and Bosker, 2015, p. 58). For example, if the regression coefficient of the within-effect for the item "Conflict between strong desires" is −0.02, it means that a within-person difference of one unit in ratings of the level of agreement with the statements "Addiction to [addiction type] is/represents a conflict between strong desires" gives a 0.02 unit decrease in the ratings of responsibility. This effect is based on the variation within the individuals, in their ratings of the nine different addiction types for the conflict-between-desires items, after controlling for mean ratings of the nine addiction types. If the regression coefficient of the between-effect is 0.06, it means that a participant with an average level of agreement of 5.5 on the items concerning "Conflict between strong desires" typically rate judgments of responsibility 0.06 points higher than a participant with an average rating of 4.5. Thus, the between-participant effect can be positive even if the within-participant effect is negative. The within-effect can be considered as similar to the type of coefficient one would obtain with so-called fixed-effect (FE) models used by economists

TABLE 1 | Unstandardized regression coefficients from 13 regression analyses predicting responsibility judgments from endorsement of addiction descriptions, sorted from negative to positive on the within-subject effects, Study 1.

	Within	SE	Between	SE	*N*	Obs.
Ordinary disease	−0.07*	0.010	−0.12*	0.019	1993	16,478
Mental disorder	−0.05*	0.008	−0.04	0.017	1979	15,855
Conflict between strong desires	−0.02	0.009	0.06*	0.020	1933	14,813
Compulsive behavior	−0.00	0.008	0.10*	0.021	1994	15,985
Reduced self-determination	0.00	0.008	0.12*	0.020	1996	16,347
Strong urge	0.01	0.010	0.10*	0.027	2008	16,586
Strong appetite	0.01	0.009	0.03	0.015	1853	13,444
Craving	0.02	0.010	0.12*	0.030	2011	16,595
Obsession	0.02	0.009	0.14*	0.023	2004	16,528
Reduced rationality	0.02*	0.008	0.23*	0.023	2008	16,601
Reduced willpower	0.04*	0.009	0.20*	0.022	2000	16,355
Reduced morality	0.05*	0.008	0.18*	0.019	1989	16,167
Habit	0.06*	0.009	0.10*	0.022	2008	16,439

*Obs = number of observations. Control variables in the regressions: addiction type (nine categories) and questionnaire version. *p < 0.01.*

(see Bell and Jones, 2015), and the between-effect approximately represents the effect one would obtain if we for each participant aggregated his/her nine ratings of agreement with a given statement and used this aggregated score as a predictor of the participants' average level of responsibility judgments. In addition to the above within-individual, between-individual, and addiction type variables, the regressions included subjects (ID variable for each respondent) as random intercepts and questionnaire version[1] as a dummy-coded, fixed-effect control variable. We used a threshold of $p < 0.01$ to identify the most promising effects in Study 1.

Results and Discussion

Table 1 presents the results of the 13 separate analyses of the addiction descriptors, ranked by the strength and direction of association with the responsibility measure. If we focus on the within-subject effects, as these adjust for time-invariant confounders (such as a general tendency to agree/disagree with survey questions, or the main effects of respondents' backgrounds), we found that endorsement of the descriptors "reduced rationality", "reduced willpower", "reduced moral competence", and "habit" were all associated with a higher level of responsibility ratings, while "ordinary disease" and "mental disorder" were negatively related to responsibility ratings.

Thus, we identified several descriptors that were associated with responsibility, notably those referring to disease or disorder, and those related to reduced ability to make the right decisions (reduced rationality, reduced morality) or control impulses (reduced willpower and habit). The descriptors conceptualizing addition as strong motivation (urges, appetites, cravings, and obsessions) were not related to judgments of responsibility in terms of within-person effects. However, we found generally stronger associations between judgments of responsibility and endorsements between individuals than within individuals. We do not have a definite explanation for this, but it might be, for example, that people's general conceptualizations of addiction matter more for responsibility judgments than do perceived differences between addiction types, or that the larger between-person effects simply reflect omitted variables related to participants' characteristics.

Study 2

Although we believed that the results from Study 1 hinted at a causal link from addiction conceptualizations to responsibility judgments, other reasons might explain the covariation. In Study 2, we wished to directly probe whether different descriptions implied different responsibility judgments by asking respondents to accept different conceptualizations and then judge the moral responsibility for addiction. For this purpose, we selected the most promising addiction descriptors from Study 1, that is, the descriptors that appeared to have within-person effects. The within-person effects are not confounded by stable characteristics of the participants (e.g., if younger participant were less familiar with the term "habit" and also more lenient in terms of ratings of responsibility, this would give a positive correlation between the two). In addition, we included two other descriptors that were not among the items in the large survey used in Study 1. The descriptor "brain disease" was included because it has become commonplace to define addiction as a chronic, relapsing brain disease (e.g., Leshner, 1997), and the same is the case for the label "irresistible desire" (Morse, 2004; Foddy, 2011).

The purpose of Study 2 was to identify conceptualizations of addiction that entailed higher or lower attributions of responsibility. Instead of exploiting existing natural variations, we asked about moral responsibility under different conceptualizations of addiction. We adjusted the wording of the responsibility question to underline the *moral* dimension of responsibility, and instead of responsibility for *becoming* addicted, we asked about the moral responsibility of *being* addicted, as these may differ (see e.g., Weiner et al., 1988).

Method

Forty-five respondents living in the United States were recruited from Amazon Mechanical Turk (Mturk). We did not collect any demographic information (but see e.g., Difallah et al., 2018, for typical characteristics of Mturk respondents).

In this within-subject study, the questions had the following format: "Given that drug addiction is [descriptor], to what extent are addicts morally responsible for their addiction?" Participants were asked to make judgments of moral responsibility for eight descriptors. The descriptors are presented in **Table 2**. Moral responsibility was measured on a scale from "not responsible at all" (0) to "fully responsible" (5). We did not specify the type of drug addiction. After responding to the eight descriptors, participants completed another set of survey items. The results of these are not reported here but were used in power calculations for Study 3.

Results and Discussion

Table 2 shows the mean levels of moral responsibility ratings for the various descriptive labels of addiction, sorted from low to high levels of responsibility. The lowest moral responsibility ratings were made when addiction was defined as a disease or a disorder, and the highest moral responsibility ratings were

TABLE 2 | Mean ratings of moral responsibility for eight descriptors, sorted from low to high, Study 2.

Given that drug addiction is…	*M* (SD)
…a mental disorder	2.51 (1.44)
…a brain disease	2.76 (1.56)
…an ordinary disease	2.93 (1.52)
…an irresistible desire	2.98 (1.56)
…a form of reduced rationality	3.00 (1.40)
…a form of reduced moral competence	3.29 (1.44)
…a strong habit	3.44 (1.23)
…a form of reduced willpower	3.58 (1.14)

[1]Half the sample received a version of the questionnaire in which they were generally asked to think about what it means to be addicted; the other half of the sample was instructed to imagine that a person close to them was addicted to a given substance. This variable was only used as a covariate in the present study and had no substantial impact on responsibility judgments.

made when descriptions directed attention toward reduced moral and rational capacities, and reduced willingness to control impulses. The moral responsibility rating of addiction as an irresistible desire was in the middle, which resulted in a pattern very similar to the ordering from diseases, *via* motivations, to reduced capacities in Study 1. A repeated-measures ANOVA obtained a $p < 0.0001$, F (4.4, 185.1) = 8.22, Eta squared = 0.16, for the test of any differences between the ratings. The results suggest that if one succeeds in changing the way addiction is represented, this could potentially influence judgments of moral responsibility.

As in Study 1, the differences in ratings appeared to reflect a continuum from uncontrollable states to reduced capacity to choose, which is consistent with the ideas from the literature on addiction models, attribution theory, and the psychological model of free will, as presented in the introductory sections. However, it is noteworthy that the mean ratings fell within a rather narrow interval at the higher end of the scale (2.5–3.6 on a scale from 0 to 5, all medians and modes were either 3 or 4). This suggests that lay perceptions may have more in common with recent models of addiction (Henden et al., 2013; Holton and Berridge, 2013; Heather, 2017a) than with a pure choice model or a pure disease model.

Study 3

In Study 3, we extended the conceptualization of addiction beyond the use of simple addiction labels by providing more detailed information about processes underlying addiction. Although the label "mental disorder" received the lowest ratings of moral responsibility in the previous studies, we believed that the brain disease conceptualization would be more relevant in terms of contemporary debates, and perhaps also easier to alter through provision of information. Scientists are increasingly discovering more about the neural mechanisms underlying addiction, and accumulated evidence suggests that repeated drug use leads to long-lasting changes in the brain. According to the brain disease model of addiction, these changes result in hijacking of the brain's reward system, impairing the autonomy and restricting addicted persons' ability to abstain from drugs, frequently denoted compulsive use (Henden et al., 2013; Pickard, 2017a,b). The modern lay person will be increasingly exposed to such reductive, mechanistic behavioral explanations couched in the neuroscientific language of neural and chemical processes. The slogan "my brain made me do it" has already become a salient feature of media, and people tend to ascribe free will and moral responsibility only to agents whose actions can be understood in terms of mental states (i.e., beliefs, desires, and intentions; Nahmias et al., 2007; De Brigard et al., 2009). Accordingly, we exposed one group of participants to detailed descriptions of brain mechanisms related to repeated drug use to see whether this could decrease their perception of the level of moral responsibility in comparison to a control group who received no particular information regarding addiction. Although we carried out the present data collection before Racine et al. (2017) published their study, our study is partly a conceptual replication of their text-only condition, for which they did not find a statistically significant effect on their free will responsibility scale.

Nahmias and Murray (2010) have noted that if one provides lay people with more concrete information about specific persons performing specific actions in specific circumstances, people engage their mind-reading abilities and consider the beliefs, desires, and intentions of agents, and thus more likely evoke judgments of free will and moral responsibility. Based on this idea, we exposed another group of participants to information depicting addiction as a brain disease and information about concrete actions needed to satisfy the addiction. We believed this focus on concrete behavior would invoke ideas about the agent's intentions and therefore undo or counteract the potential impact of the neuromechanistic information.

The descriptors in Studies 1 and 2 that resulted in the lowest ratings of moral responsibility represented states and physical conditions of the individual, whereas the labels with the highest ascription of responsibility concerned behavior or capacities relating to behaviors (e.g., habit and willpower). One could argue that addiction as a state connotes elements of inaction, directing attention toward identity and a definition of someone as a certain kind of person. Having a status or identity does not necessarily mean that one acts out one's identity. This point led us to investigate whether framing addiction as a state (being addicted) or as a behavior (performing actions to satisfy addiction) by changing the object of evaluation could influence judgments of moral responsibility. In summary, Study 3 tested two different ways of altering addiction conceptualizations: provision of information and framing addiction in terms of a behavior or a state.

Method

Sample and Design

Based on results from pilot data[2], we chose the sample size such that it would give 80% power for a one-tailed test with a p- threshold of 0.05 for the comparison between the addiction state framing versus addictive behavior framing. Data from 1,062 Mturk participants were collected. The full design was a 2 (addiction type; within-person) by 3 (addiction information; between-person) by 2 (addiction framing; between-person) experimental design. Confidence intervals of mean differences were based on estimated marginal means from a repeated-measures ANOVA in SPSS 24.

Experimental Conditions and Measures

No Information

In the information control condition, participants did not get any information about addiction before they made responsibility judgments.

[2]The pilot data were from the same survey as the data reported in Study 2. The difference between a state vs. behavior condition in the pilot data was 0.2 points on a scale from 0 to 5, 95% CI [0.047–0.377], $F(1, 43) = 6.70$, $p = 0.013$. We later discovered that Albers and Lakens (2018) recommended *not* to calculate power directly based on pilot data effect sizes.

Brain Disease Information

In the brain disease information condition, participants received the following information, based on various internet resources:

> In recent years, more and more research suggests that drug addiction can be viewed as a form of brain disease. The following text is based on information from the web page of the American Society of Addiction Medicine:
>
> Research shows that the brain disease of addiction affects neurotransmission and interactions within the reward circuitry of the brain so that addictive behaviors substitute for normal healthy behaviors, and memories of previous experience with drugs trigger craving and desire for more addictive behavior. The disease creates distortions in thinking, feelings and perceptions. Addictive behaviors are manifestations of the brain disease, and the final result is a dysfunctional pursuit of rewards when seeking more drugs.
>
> Here is another excerpt from a neuroscientist:
>
> All drugs of abuse, from nicotine to heroin, provide a release of dopamine that creates a feeling of pleasure. In addition, this release of dopamine affects learning and memory. Addictive substances stimulate the same circuit in the brain that becomes activated by natural rewards such as sex and food. However, drugs overstimulate the circuit and the reward system responds with less production of dopamine—an adaptation similar to turning the volume down on a loudspeaker when noise becomes too loud. As a result of these adaptations, dopamine has less impact on the brain's reward center so that the desired substance no longer gives as much pleasure as before. Addicts have to take more of the drug to obtain the same dopamine "high" because their brains have adapted—an effect known as tolerance. Now compulsion takes over, a reflection of how the normal machinery of motivation is no longer functioning.

Brain Disease + Agency Information

Participants in the brain disease + agency information condition first read the same information as in the Brain disease information condition; then, they received the following text:

> An addiction to heroin typically requires planning and effort, for instance, planning how to obtain money, seeking a dealer, negotiating price, and preparing the drug before finally injecting or smoking it. An addiction to nicotine also requires planning and effort. Smokers addicted to nicotine have to buy cigarettes or tobacco, bring the cigarettes and perhaps a lighter or matches along when going out, find an appropriate place to smoke, and sometimes make plans about how to take smoking breaks that do not interfere with work or other activities.

Addiction States Versus Addictive Behavior

Orthogonal to the above three information conditions, approximately half of the respondents received the two questions "To what extent is a heroin user morally responsible for being addicted to heroin?" and "To what extent is a cigarette smoker morally responsible for being addicted to nicotine?". This was the addiction as state condition. The other half received the two questions "To what extent is a heroin user morally responsible for actions performed to satisfy the addiction to heroin?" and "To what extent is a cigarette smoker morally responsible for actions performed to satisfy the addiction to nicotine?" This was the addiction as behavior condition. The responses were recorded on a scale from 0 ("not responsible at all") to 5 ("fully responsible").

Results and Discussion

Table 3 presents the mean levels of responsibility ratings for all conditions. An ANOVA suggested that the ratings varied between the information conditions (no information, brain, and rain + agency), $F(2, 1{,}056) = 11.30$, $p < 0.0001$, Partial Eta Squared = 0.02. The information describing addiction as a brain disease in a mechanistic and reductionistic language produced lower levels of moral responsibility than did the control condition, difference = −0.44, 95% CI [−0.26, −0.62]. When the brain disease description was followed by information about the plans and concrete actions addicted persons will have to make to satisfy their addiction, the moral responsibility ratings increased somewhat in comparison with the brain description only, difference = 0.19, 95% CI [0.00, 0.37]. However, these participants were still substantially more lenient than

TABLE 3 | Ratings of moral responsibility for heroin and cigarette addiction in three information conditions by two framing conditions, Study 3.

	Addiction as state			Addiction as behavior		
		Heroin	Cigarettes		Heroin	Cigarettes
	n	*M* (SD)	*M* (SD)	*n*	*M* (SD)	*M* (SD)
No information	181	3.35 (1.35)	3.75 (1.26)	173	3.76 (1.29)	4.06 (1.21)
Brain	172	2.98 (1.29)	3.18 (1.29)	204	3.35 (1.26)	3.67 (1.11)
Brain + agency	180	3.21 (1.40)	3.43 (1.39)	152	3.54 (1.34)	3.75 (1.28)

were those in the control group, difference = −0.25, 95% CI [−0.06, −0.44]. Thus, reminding people about the intentions, plans, and concrete actions involved in sustaining an addiction (i.e. *agency*) did not appear to cancel out the effect of conceptualizing addiction at the level of neural mechanisms.

When the object of evaluation was addictive behaviors, the ratings were 0.37 (one-sided 95% CI [0.25, inf.]) points higher than when the object of judgment was addictive states, F (1, 1,058) = 24.56, $p < 0.0001$, Partial Eta Squared = 0.02. Thus, lay people considered addicted persons to be more morally responsible for actions performed to satisfy an addiction than for the state of being addicted. This result is consistent with the idea that information about agents performing specific actions should evoke perceptions of free will and moral responsibility (Nahmias and Murray, 2010). Note that in principle, if people endorse a brain disease conceptualization and accept the mechanistic brain model of addiction, the responsibility for addictive states and addictive actions should be equally low.

Although the effect sizes were small, the data clearly showed that it is possible to manipulate people's immediate judgments of responsibility for addictions. This suggests that those who provide information and have the power to frame questions about addiction, like the media and professionals in the justice and health care systems, also have the power to change people's moral judgments about addicted individuals.

GENERAL DISCUSSION

In the present studies, we investigated the relation between conceptualizations of addiction and moral responsibility. To our knowledge, this is the first study where various labels and descriptions from the addiction literature are mapped onto a dimension of lay moral responsibility. Furthermore, the study showed that lay people's moral judgments were malleable, which in past studies have been difficult to demonstrate.

Correlational data in Study 1 indicated that endorsement of labels that described addiction as a disease or disorder was associated with lower ratings of responsibility, whereas endorsement of labels relating to behavior and choice was associated with higher ratings of moral responsibility. This pattern was confirmed in Study 2 when lay people were asked to adopt certain perspectives and asked to make judgments about moral responsibility.

In Study 3, we observed that providing detailed information about brain mechanisms and neural changes following drug intake lowered ratings of moral responsibility. Adding information about the behaviors needed to satisfy an addiction reduced the effect of the brain mechanism information but did not fully cancel out the effect. A similar pattern of more responsibility for actions was also found when we manipulated the object of evaluation. Participants attributed more responsibility to addictive actions than addictive states.

In general, the studies demonstrate that conceptualizations of addictions can be consequential for judgments of moral responsibility. This may not seem to align with the findings of past research (Rather, 1991; Meurk et al., 2014; Racine et al., 2017). However, the study by Racine et al. (2017) showed similar tendencies as in our studies, and they noted that a more strongly worded message may be more successful in changing how people view addiction. Furthermore, past research has already documented that different conceptualizations, in the form of perceptions of different types of addictions, are consequential for moral judgments (Blomqvist, 2009; Rise et al., 2014). These past results on different types of addictions could be due to numerous factors, such as how common the addictive behavior is, how often people quit, how serious the health consequences are, what kind of people are associated with the behavior, and so on. In the present research, we either controlled for the average effect of the specific behavior (Study 1), or we manipulated conceptualizations while holding the behaviors constant (Studies 2 and 3). Thus, the present study shows how conceptualizations of addiction, irrespective of the nature of the specific addictive behavior, affect moral judgments.

The three studies used very different methods, from asking participants to rate how well a label represents addiction, to changing the object of evaluation. Still, we assume that the results reflect the same phenomenon, namely how the flexibility of people's views of addiction can produce differences in their judgments of its moral consequences. People hold different views on different types of addictions, and this appears to be consequential for their moral judgments (Study 1). People are able to quickly change their inferences regarding moral consequences of addiction when we ask them to link addiction to other known concepts such as disease and habit (Study 2). People's judgments of moral responsibility change when we provide new information or information that remind them of certain aspects of addiction, and people's judgments change when we frame addiction as a behavior instead of a state (Study 3). Interestingly, regardless of the way addiction is malleable (i.e., addiction type, link to other concepts, provision of information, and framing of addiction), the relationship between the conceptualizations and their consequences for moral judgments appear to follow a predictable pattern, which is discussed below.

ADDICTION MODELS, ATTRIBUTION, AND FREE WILL

Theoretically, the present results resonate well with the ideas described in the introductory sections. The explorative analyses of addiction labels in Studies 1 and 2 revealed that states and disease models of addiction were associated with lower levels of responsibility and labels implying reduced choice capacity or self-control failure were associated with higher levels of responsibility. Lay people's intuitive judgments lie somewhere in the middle of the two extreme poles of moral responsibility, with only slight variation, depending on whether addiction is conceptualized as disease or choice/behavior. The placement of addiction in the middle of a moral responsibility continuum is consistent with recent models of addiction (Henden et al., 2013; Holton and Berridge, 2013; Heather, 2017a).

The results were also consistent with the idea that moral judgments are based on perceptions of controllability of cause

(e.g., Weiner, 1995) or perceptions of intent (Shaver, 1985). Presumably, people think that concrete behaviors are controllable, whereas being addicted is not so controllable. This was particularly clear in Study 3, where we manipulated the object of evaluation, and observed higher ratings of responsibility for addictive behavior than for being addicted. Furthermore, the specific information about actions given after the neuromechanistic information also pointed to the potential of intentional control over addiction, and it appeared to reduce some of the effect of the neuromechanistic information.

In the introductory sections, we presented a psychological lay model of free will as degrees of agency. Addicts are agents who have capacities to decide and exercise control but who are also subject to internal and external constraints (cf. Nahmias, 2018). Most likely, lay people know that addicts have a strong desire for the drug, experience a lot of psychological distress and that they may not have many available alternative courses of actions. In other words, people may perceive addicts as having free will but not being fully free agents. Judged by the pattern of moral judgments in the present studies, the notion that addicts have free will and at the same time are unfree agents does not seem to represent a paradox for lay people. This also seemed to be the case in the study by Wiens and Walker (2015), where adopting a disease model did not have any impact on beliefs in free will but still reduced beliefs in agency.

Similarly, it appears that lay people see no contradiction in thinking of addicts as simultaneously *intentional* agents and unfree agents. Reminding lay people that consumption of a drug requires an elaborate series of planning, preparation, and effortful actions in advance of consumption in Study 3 (i.e., addicts are in effect agents with an intact intentional system) did not lead them to fully ignore the brain information. Perhaps the research participants were thinking that the elaborate efforts to satisfy the addiction could be propelled by a strong desire, thus bypassing the intentional system.

Even when asked to accept a mechanistic disease view, lay people were more willing to attribute moral responsibility for addictive actions than for states. This may reflect the perception that addicted persons have a choice when performing concrete actions but still have an underlying condition that limits agency and serves as an excuse for *being* addicted. This pattern of judgments suggests that lay people hold a model similar to the *disorder of choice model* advocated by Heather (2017b): "[…]what is needed is a model that continues to see addiction as behavior that people find extremely difficult to change while at the same time accepting the obvious fact of voluntary drug-seeking and – taking." Although lay people's perception of agency decreases when addiction is described as a disease, they still consider addicted individuals to be moral agents with a capacity for choice.

IMPLICATIONS

This study among lay people provides evidence that conceptualizations of addiction matter for assignment of moral responsibility, with addictive labels related to choices and behaviors increasing the level of moral responsibility and labels related to brain disease lowering the level of responsibility. Based upon the present data it may, in principle, be possible to raise or lower the level of moral responsibility by manipulating the description of addiction. If one wishes a high level of moral responsibility for addiction, one could provide a minimal amount of information about the etiology and mechanisms of addiction and focus upon addictive actions. In contrast, if one wants a low level of moral responsibility, one could conceptualize addiction as a disease or disorder by providing information about brain mechanisms or using labels relating to disease and mental disorders. Motivational labels, like urge and desire, may be more neutral (Study 1) or imply an intermediate level of responsibility (Study 2).

The labels and descriptions used by, for example, the media and scientists might influence the public and policy makers, and, in turn, affect how addicted individuals are treated. Lay perceptions of moral responsibility of addictions have been shown to be a predictor of how much help addicted individuals deserve in the sense that higher levels of moral responsibility lowers the level of deservingness of help (e.g., Rise et al., 2014) and may thus function as a legitimation for policy decisions. We would be happy to see future research on consequences of moral responsibility of different conceptualizations of addiction for real life outcomes such as social interactions and policy decisions.

We do have to keep in mind that less responsibility is not necessarily beneficial for addicted individuals. Wiens and Walker (2015) found that people with a mild to moderate alcohol addiction experienced less control in relation to their drinking after being manipulated to adopt a disease model of addiction. Adopting a disease model did not reduce feelings of stigma more than adopting a psychosocial model. Similarly, it has been shown that lay models of psychiatric disorders based on biological mechanisms can increase pessimism about recovery and may increase the perception that people with psychological problems are dangerous (see Kvaale et al., 2013). Furthermore, a study by Kingree et al. (1999) suggested better outcomes in a 12-step program when participants felt more personally responsible for their addictions. On the other hand, one study showed that people who were informed that they had a genetic predisposition to alcoholism were more willing to sign up for a workshop on responsible drinking (Dar-Nimrod et al., 2013).

LIMITATIONS, STRENGTHS, AND CONCLUSIONS

In cross-sectional studies, the relation between addiction conceptualizations and responsibility judgments can be confounded by variables relating to demographics and ideology (e.g., elder people could give higher endorsement due to familiarity with the concept and also be generally more punitive). This is not a problem in the present studies as we focused on within-person effects and used experimental

manipulations. However, Study 1 did not give us any information about the direction of potential causal relations, and Study 2 only indicated consequences for responsibility judgments *given that* a certain conceptualization was accepted.

We used a direct measure of moral responsibility that is assumed to capture the process of assigning moral responsibility to events and behaviors (Weiner, 1995). Although it seems to be a common practice in experimental philosophy to use one-item measures for moral responsibility (see Cova and Kitano, 2014), this might be perceived as problematic in terms of measurement reliability. However, if we were to combine the two response measures (heroin and cigarette smoking) in Study 3 to an index, Cronbach's alpha would be as high as 0.9.

The choice of ratings of moral responsibility as our only outcome measure limits our knowledge about specific real-life consequences of adopting different addiction models. Responsibility is a rather abstract concept believed to contribute to a range of outcomes (e.g., Weiner, 1995, 2006; Halkjelsvik and Rise, 2014).

The present studies showed that natural variations in conceptualizations of addiction may be consequential for judgments of moral responsibility, that different conceptualizations imply different moral judgments, and that conceptualizations are malleable through information and through changing the focus of evaluation (states versus actions). This means that the way people describe, teach about, and frame addiction could have implications for a range of behaviors that are based on moral judgments.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

As the study involved anonymous data, voluntary participation, and did not ask for any information about participants' health or any other sensitive data, the study did not require approval from an ethics committee according to Norwegian law. Participants were informed that their responses would be used in research and asked to consent by proceeding to the questionnaire.

AUTHOR CONTRIBUTIONS

The authors contributed equally to the article. JR drafted the introductory sections and the Discussion. TH performed statistical analyses and drafted the method and results sections. Both authors designed the studies, revised the manuscript, and approved the final version.

FUNDING

The study was funded by the Norwegian Institute of Public Health.

REFERENCES

Albers, C., and Lakens, D. (2018). When power analyses based on pilot data are biased: inaccurate effect size estimators and follow-up bias. *J. Exp. Soc. Psychol.* 74, 187–195. doi: 10.1016/j.jesp.2017.09.004

Bell, A., and Jones, K. (2015). Explaining fixed effects: random effects modeling of time-series cross-sectional and panel-data. *Polit. Sci. Res. Methods* 3, 133–153. doi: 10.1017/psrm.2014.7

Berridge, V., Mold, A., Beccaria, F., Eisenbach-Stangl, I., Hercyńska, G., Moskalewicz, J., et al. (2014). Addiction in Europe, 1860s-1960s: concepts and responses in Italy, Poland, Austria, and the United Kingdom. *Contemp. Drug Probl.* 41, 551–566. doi: 10.1177/0091450914567119

Blomqvist, J. (2009). What is the worst thing you could get hooked on? *Nordic Stud. Alcohol Drugs* 26, 373–398.

Brickman, P., Rabinowitz, V. C., Karuza, J., Coates, D., Cohn, E., and Kidder, L. (1982). Models of helping and coping. *Am. Psychol.* 37, 368–384. doi: 10.1037/0003-066X.37.4.368

Corrigan, P. W., Edwards, A. B., Green, A., Diwan, S. L., and Penn, D. L. (2001). Prejudice, social distance, and familiarity with mental illness. *Schizophr. Bull.* 27, 219–225. doi: 10.1093/oxfordjournals.schbul.a006868

Corrigan, P. W., Kuwabara, S. A., and O'Shaughnessy, J. (2009). The public stigma of mental illness and drug addiction: findings from a stratified random sample. *J. Soc. Work* 9, 139–147. doi: 10.1177/1468017308101818

Corrigan, P. W., Rowan, D., Green, A., Lundin, R., River, P., Uphoff-Wasowski, K., et al. (2002). Challenging two mental stigmas: personal responsibility and dangerousness. *Schizophr. Bull.* 28, 293–309. doi: 10.1093/oxfordjournals.schbul.a006939

Cova, F., and Kitano, Y. (2014). Experimental philosophy and the compatibility of free will and determinism: a survey. *Ann. Jpn. Assoc. Philos. Sci.* 22, 17–37. doi: 10.4288/jafpos.22.0_17

Dar-Nimrod, I., Zuckerman, M., and Duberstein, P. R. (2013). The effects of learning about one's own genetic susceptibility to alcoholism: a randomized experiment. *Genet. Med.* 15, 132–138. doi: 10.1038/gim.2012.111

De Brigard, F., Mandelbaum, E., and Ripley, D. (2009). Responsibility and the brain sciences. *Ethic. Theory Moral Prac.* 12, 511–526. doi: 10.1007/s10677-008-9143-5

Difallah, D., Filatova, E., and Ipeirotis, P. (2018). "Demographics and dynamics of mechanical Turk workers" in *Proceedings of WSDM 2018: The eleventh ACM international conference on web search and data mining*. (Marina Del Rey, CA, USA: ACM, New York), February 5–9, WSDM 2018.

Feldman, G., Baumeister, R. F., and Wong, K. F. E. (2014). Free will is about choosing: the link between choice and the belief in free will. *J. Exp. Soc. Psychol.* 55, 239–245. doi: 10.1016/j.jesp.2014.07.012

Foddy, B. (2011). Addiction and its sciences—philosophy. *Addiction* 106, 25–31. doi: 10.1111/j.1360-0443.2010.03158.x

Halkjelsvik, T., and Rise, J. (2014). Social dominance orientation, right wing autortarianism and willingness to help addicted individuals: the role of responsibility judgment. *Eur. J. Psychol.* 10, 27–40. doi: 10.5964/ejop.v10i1.669

Heather, N. (2017a). "Overview of addiction as a disorder of choice and future prospects" in *Addiction and chioce. Rethinking the relationship*. eds. N. Heather and G. Segal (Oxford, UK: Oxford University Press), 463–482.

Heather, N. (2017b). Q: is addiction a brian disease or a moral failing? A: neither. *Neuroethics* 10, 115–124. doi: 10.1007/s12152-016-9289-0

Heather, N., Best, D., Kawalek, A., Field, M., Lewis, M., Rotgers, F., et al. (2017). Challenging the brain disease model of addiction: European launch of the addiction theory network (Editorial). *Addict. Res. Theory* 25, 1–7. doi: 10.1080/16066359.2017.1399659

Henden, E., Melberg, H. O., and Røgeberg, O. (2013). Addiction: choice or compulsion? *Front. Psych.* 4:141. doi: 10.3389/fpsyt.2013.00077

Heyman, G. M. (2009). *Addiction: A disorder of choice*. (Cambridge: Harvard University Press).

Holton, R., and Berridge, K. (2013). "Addiction between compulsion and choice" in *Addiction and self-control. Perspectives from philosophy, psychology, and neuroscience*. ed. N. Levy (Oxford: Oxford University Press), 239–268.

Kennett, J., and McConnell, D. (2013). Explaining addiction: how far does the reward account of motivation take us? *Inquiry* 56, 470–489. doi: 10.1080/0020174X.2013.806133

Kingree, J. B., Sullivan, B. F., and Thompson, M. P. (1999). Attributions for the development of substance addiction among participants in a 12-Step oriented treatment program. *J. Psychoactive Drugs* 31, 129–135. doi: 10.1080/02791072.1999.10471735

Kvaale, E. P., Gottdiener, W. H., and Haslam, N. (2013). Biogenetic explanations and stigma: a meta-analytic review of associations among laypeople. *Soc. Sci. Med.* 96, 95–103. doi: 10.1016/j.socscimed.2013.07.017

Leshner, A. (1997). Addiction is a brain disease, and it matters. *Science* 278, 45–47. doi: 10.1126/science.278.5335.45

Levy, N. (2011). "Addiction, responsibility, and ego-depletion" in *Addiction and responsibility*. eds. J. Poland and G. Graham (Cambridge: MIT Press), 89–111.

Lewis, M. (2015). *The biology of desire: Why addiction is not a disease*. (New York: Perseus Books Group).

Melberg, H. O., Henden, E., and Gjelsvik, O. (2013). Addiction and responsibility: a survey of opinions. *Inquiry* 56, 558–570. doi: 10.1080/0020174X.2013.806143

Mele, A. R. (2006). *Free will and luck*. (Oxford; New York: Oxford University Press).

Meurk, C., Carter, A., Partridge, B., Lucke, J., and Hall, W. (2014). How is acceptance of the brain disease model of addiction related to Australians' attitudes towards addicted individuals and treatments for addiction? *BMC Psychiatry* 14:373. doi: 10.1186/s12888-014-0373-x

Monroe, A. E., Brady, G., and Malle, B. F. (2016). This isn't the free will worth looking for: general free will beliefs do not influence moral judgements, agent-specific choice ascriptions do. *Soc. Psychol. Personal. Serv.* 8, 191–199. doi: 10.1177/1948550616667616

Monroe, A. E., and Malle, B. F. (2010). From uncaused will to conscious choice: the need to study, not speculate about people's folk concept of free will. *Rev. Philos. Psychol.* 1, 211–224. doi: 10.1007/s13164-009-0010-7

Monroe, A. E., and Malle, B. F. (2015). "Free will without metaphysics" in *Surrounding free will*. ed. A. Mele (New York, NY: Oxford University press).

Morse, S. J. (2004). Medicine and morals, craving and compulsion. *Subst. Use Misuse* 39, 437–460. doi: 10.1081/JA-120029985

Nahmias, E. (2018). "Free will as a psychological accomplishment" in *Oxford handbook on freedom*. eds. D. Scmidtz and C. Pavel (New York: Oxford University Press), 492–507.

Nahmias, E., Coates, D. J., and Kvaran, T. (2007). Free will, moral responsibility, and mechanism: experiements on folk intuitions. *Midwest Stud. Philos.* 31, 214–242. doi: 10.1111/j.1475-4975.2007.00158.x

Nahmias, E., and Murray, D. (2010). "Experimental philosophy on free will: an error theory for incompatibilist intuitions" in *New waves in philosophy of action*. eds. J. Aguliar, A. Buckareff, and K. Frankish (New York: Palgrave-MacMillan), 189–215.

Nichols, S., and Knobe, J. (2007). Moral responsibility and determinism: the cognitive science of folk intuitions. *Noûs* 41, 663–685. doi: 10.1111/j.1468-0068.2007.00666.x

Pickard, H. (2017a). "Addiction" in *The Routledge companion to free will*. eds. M. Griffith, N. Levy, and K. Timpe (New York: Routledge), 454–467.

Pickard, H. (2017b). Responsibility without blame for addiction. *Neuroethics* 10, 169–180. doi: 10.1007/s12152-016-9295-2

Pickard, H., and Pierce, S. (2013). "Addiction in context: philosophical lessons from a personality disorder clinic" in *Addiction and self-control. Perspectives from philosophy, psychology, and neuroscience*. ed. N. Levy (Oxford: Oxford University Press), 165–189.

Pinfold, V., Toulmin, H., Thornicraft, G., Huxley, P., Farmer, P., and Graham, T. (2003). Reducing psychiatric stigma and discrimination: evaluation of educational interventions in UK secondary schools. *Br. J. Psychiatry* 182, 342–346. doi: 10.1192/bjp.182.4.342

Racine, E., Sattler, S., and Escande, A. (2017). Free will and the brain disease model of addiction: the not so seductive allure of neuroscince and its modest impact on the attribution of free will to people with an addiction. *Front. Psychol.* 8:1850. doi: 10.3389/fpsyg.2017.01850

Rather, B. C. (1991). Disease versus social learning models of alcoholism in the prediction of alcohol problem recognition, help seeking, and stigma. *J. Drug Educ.* 21, 119–132.

Rise, J., Aarø, L. E., Halkjelsvik, T., and Kovac, V. B. (2014). The distribution and role of causal beliefs, inferences of responsibility, and moral emotions on willingness to help addicts among Norwegian adults. *Addict. Res. Theory* 22, 117–125. doi: 10.3109/16066359.2013.785532

Rise, J., Halkjelsvik, T. B., and Kovac, V. B. (2015). Mental states of addiction: conceptions in the adult population. *Contemp. Drug Probl.* 42, 289–298. doi: 10.1177/0091450915608446

Room, R., Hellman, M., and Stenius, K. (2015). The dance between concepts and terms. *Intern. J. Alcohol Drug Res.* 4, 27–35. doi: 10.7895/ijadr.v4i1.199

Sattler, S., Escande, A., Racine, E., and Gŏritz, A. S. (2017). Public stigma toward people with drug addiction: a factorial study. *J. Stud. Alcohol Drugs* 78, 415–425. doi: 10.15288/jsad.2017.78.415

Schulze, B., Richter-Werling, M., Matschinger, H., and Angermeyer, M. C. (2003). Crazy? So What! Effects of a school-based project on students' attitudes towards people with schizophrenia. *Acta Psychiatr. Scand.* 107, 142–153. doi: 10.1034/j.1600-0447.2003.02444.x

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. (New York: Springer-Verlag).

Sinnott-Armstrong, W. (2013). "Are addicts responsible?" in *Addiction and self-control: Perspectives from philosophy, psychology, and neuroscience*. ed. N. Levy (Oxford: Oxford University Press), 122–143.

Snijders, T. A. B., and Bosker, R. J. (2015). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. 2nd edn. (London: Sage Publishers).

Uusitalo, S. (2011). On addicts' moral responsibility and action. *Res. Cogitans* 1, 77–91. doi: 10.2478/nsad-2013-0004

Uusitalo, S., Salmela, M., and Nikkinen, J. (2013). Addiction, agency and affects–philosophical perspectives. *Nordic Stud. Alcohol Drugs* 30, 33–50. doi: 10.2478/nsad-2013-0004

Vonasch, A. J., Baumeister, R. F., and Mele, A. R. (2018). Ordinary people think free will is a lack of constraint, not the presence of a soul. *Conscious. Cogn.* 60, 133–151. doi: 10.1016/j.concog.2018.03.002

Vonasch, A. J., Clark, C. J., Lau, S., Vohs, K. D., and Baumeister, R. F. (2017). Ordinary people associate addiction with loss of free will. *Addict. Behav. Rep.* 5, 56–66. doi: 10.1016/j.abrep.2017.01.002

Weiner, B. (1995). *Judgment of responsibility. A foundation for a theory social conduct*. (New York: The Guilford Press).

Weiner, B. (2006). *Social motivation, justice, and the moral emotions. An attributional approach*. (London: Lawrence Erlbaum).

Weiner, B., Perry, R. P., and Magnusson, J. (1988). An attributional analysis of reactions to stigmas. *J. Pers. Soc. Psychol.* 55, 738–748. doi: 10.1037/0022-3514.55.5.738

Wiens, T. K., and Walker, L. J. (2015). The chronic diseaes concept of addiction. Helpful or harmful? *Addict. Res. Theory* 23, 309–321. doi: 10.3109/16066359.2014.987760

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

PERSPECTIVE
published: 21 January 2020
doi: 10.3389/fpsyg.2019.03009

The Art of Influencing Consumer Choices: A Reflection on Recent Advances in Decision Neuroscience

Nadège Bault[1,2]* *and Elena Rusconi*[3]

[1] School of Psychology, University of Plymouth, Plymouth, United Kingdom, [2] Center for Mind/Brain Sciences (CIMeC), University of Trento, Trento, Italy, [3] Department of Psychology and Cognitive Science, University of Trento, Trento, Italy

In recent years, our knowledge concerning the neurobiology of choice has increased tremendously. Research in the field of decision-making has identified important brain mechanisms by which a representation of the subjective value of an option is built based on previous experience, retrieved and compared to that of other available options in order to make a choice. One body of research, in particular, has focused on simple value-based choices (e.g., choices between two types of fruits) to study situations very similar to our daily life decisions as consumers. The use of neuroimaging techniques has deepened and refined our knowledge of decision processes. Additionally, computational approaches have helped identifying and describing the mechanisms underlying newly found components of the decisional process. They provide mechanistic explanations for diverse biases that can drive decision makers away from their own preferences or from rational choices. It is now clear that both attentional and affective factors can exert robust effects on an individual's decisions. Because these factors can be manipulated externally, academic research and theories are of great interest to the marketing industry. This approach is becoming increasingly effective in manipulating consumer behavior and has the potential to become even more effective in the future. Another line of research has revealed differences in the decision-making neural circuitry that underlie sub-optimal choice behavior, rendering some individuals particularly vulnerable to marketing strategies. As neuroscientists, we wonder whether relevant institutions should direct their efforts toward raising citizens' awareness, demanding more transparency on marketing applications and regulate the most pervasive communication techniques in marketing, in view of their current use and of recent research progress.

Keywords: value-based decisions, choice biases, marketing, regulation, decision neuroscience

OPEN ACCESS

Edited by:
Marco Tullio Liuzza, University of Magna Graecia, Italy

Reviewed by:
Martin Skov, Copenhagen Business School, Denmark
Cinzia Calluso, Guido Carli Free International University for Social Studies, Italy

***Correspondence:**
Nadège Bault
nadege_lab@nbault.net

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *26 August 2019*
Accepted: *19 December 2019*
Published: *21 January 2020*

Citation:
Bault N and Rusconi E (2020) The Art of Influencing Consumer Choices: A Reflection on Recent Advances in Decision Neuroscience. Front. Psychol. 10:3009. doi: 10.3389/fpsyg.2019.03009

ATTENTIONAL BIASES IN CONSUMER CHOICES

Tremendous progress has been realized in the last decade in our understanding of attentional effects on decision processes, through the description of their neurocomputational mechanisms. Thus, we will focus here on those mechanisms to illustrate how they can inform marketing strategies. Psychological and neural accounts of the role of memory and affective mechanisms in consumer decisions can be found in Plassmann and Karmarkar (2015).

Cognitive and Neural Mechanisms of Simple Choice

When facing a simple decision, for instance picking a fruit to eat in a basket containing several types of fruits, our brain computes a value signal. The value represents the expected benefit of

consuming the good based on previous experience. Recent cognitive models of decision-making propose that a value is assigned to all the options available, then the values are compared in order to reach a decision (Rangel and Clithero, 2014). Expected delays, potential price, or uncertainty in its obtainment of the good will all be incorporated into the value signal. How exactly the value is computed, though, is still under scrutiny. Much evidence supports the theory that values are computed through reinforcement learning. A value is updated when our experience in consuming the good does not match our expectations, a mechanism that supports adaptive behavior. This learning mechanism is implemented in the brain by dopaminergic neurons of the ventral striatum. These neurons encode a prediction error signal which serves as an update signal for the value (Schultz, 1998; Tobler et al., 2005). They project to a frontal region called the ventromedial prefrontal cortex, which is thought to store the value signal (Ruff and Fehr, 2014). However, the way we value options often depends on our internal states (e.g., how hungry we are at that particular moment) and on states of the world (we might value more a juicy fruit in the summer than in the winter). Assuming that values of goods are stored globally fails to explain why choices can vary with the decision context.

Another theory proposes that we separately evaluate all attributes of the available options and integrate them at the time of choice (Rangel and Clithero, 2014). The value of an apple is not represented as such; rather, value associated with its color, taste, smell or shape are encoded separately. Considering the attributes of a good and retrieving the values associated with those attributes requires attention.

The Influence of Attention on Decisions

Krajbich and Rangel (2011) proposed that attention fluctuates among the different items being evaluated during a decision, and this affects the computation of their value. They applied a well-established model in perceptual decision-making (Ratcliff, 1978; Ratcliff et al., 2016) to simple value-based choices in order to characterize the link between attention – as measured by eye gaze and decision latency – to decision output through the hidden value computation process. Their attentional drift diffusion model (aDDM), applied to binary choices, states that the values of the attributes of the currently attended item are retrieved and integrated (Krajbich et al., 2010; but see Summerfield and Tsetsos, 2012; Calluso et al., 2015; for alternative drift diffusion models of value-based decision). At any point of time, the integrated value is then compared to the value of the unattended item. The agent freely explores the available options, switching their attention among the items. If the two items are appetitive (i.e., have been associated with positive experience in the past), the retrieval of their value will yield to a positive signal. While a specific item is being fixated, its value is computed and its relative value, compared to the other item, increases. When the difference between the values of the two items reaches a given threshold, the decision process terminates (Krajbich et al., 2010).

Evidence supporting this model is provided by experiments in perceptual decision-making (e.g., is the left segment shorter than the right one?) showing that in every choice, the firing rate of neurons increases proportionally to the easiness of the decision (integration process) and reaches the same point (threshold) right before an answer is given (Roitman and Shadlen, 2002; Gold and Shadlen, 2007). Moreover, during binary choices between snacks, the striatum and the ventromedial prefrontal cortex (i.e., two brain areas involved in valuation and choices) encode the value of the attended item, relatively to the value of the unattended item (Hare et al., 2011; Lim et al., 2011). Thus, attention modulates brain activity related to the retrieval and comparison of values.

The theory has several implications which have been verified experimentally. First, because the value of a desirable item increases when it is attended, the chosen item is the last one to be fixated before the threshold is reached and the decision is made. Second, the first fixated item gets an advantage in the value computation process and thus is more likely to be chosen. Third, the longer an item is being looked at the more likely it is that it will be chosen. Using repeated choices between snacks in combination with eye tracking, Krajbich et al. (2010) were able to confirm all those predictions. When choosing between two snacks equally liked by participants, they picked the last fixated item in about 75% of the trials. Moreover, the longest the first fixation, the higher the probability that the corresponding item would be chosen. Lastly, the longest an item was fixated and the higher was the probability it would be chosen, even after correcting for liking ratings. Importantly, similar choice biases induced by fixation trajectories were observed during purchasing decisions (Krajbich et al., 2012).

Manipulating Attention to Bias Consumer Choices

As decision processes are strongly influenced by visual exploration, this evidence may imply that externally orienting attention would result in systematic decision biases. Indeed, controlling the duration of visual presentation of the options can change judgments about the attractiveness of human faces (Shimojo et al., 2003) and about moral situations (Pärnamets et al., 2015). Decisions to acquire food or art items (Armel et al., 2008; Lim et al., 2011)[1] can be biased as well. The likelihood that an item is chosen increases between 6 and 11% when it was seen for 900 ms rather than 300 ms. Therefore, people have a bias to choose the things they have been viewing the longest rather than those they genuinely prefer. Gaze patterns reflect the preferences of individuals; they influence those preferences as well.

In addition, visually salient items would grab more attention (Itti and Koch, 2001), hence be fixated first and longer, and ultimately be chosen more often. Studies have shown that manipulating the visual saliency of stimuli by varying features such as intensity, color, and orientation results in participants making a choice that contradicts their initial preferences (Navalpakkam et al., 2010; Towal et al., 2013). These effects extend to purchasing environments, where

[1] A demonstration of the effect is available in a TEDx talk delivered by Antonio Rangel (http://www.tedxcaltech.com/content/antonio-rangel).

they can become even stronger when the cognitive load is high. The color, and brightness of the packaging can lead individuals to choose their least preferred product under time pressure (Milosavljevic et al., 2012). Similarly, the probability that individuals will pick the brand they value the most in a supermarket shelf decreases as the number of available products increases. They tend to grab the product right in front of them. Because of reading habits, in occidental countries, options placed in the top left corner are chosen more often than those in lower right corner (Reutskaja et al., 2011).

Applications in Marketing

Clearly, advertisers did not wait for psychologists and neuroscientists to describe the cognitive mechanisms of the attention grabbing effects on decisions to exploit them (Pieters and Wedel, 2004). Nonetheless as academic research makes progress in identifying decision biases, precisely describing the variables that can cause these biases in more and more refined theoretical models, advertising and other marketing techniques will become more effective. In fact, many efforts are directed into bridging neuroscience research with marketing both at the academic and at the industry levels (Plassmann et al., 2007; Karmarkar and Plassmann, 2019). Marketing companies are now equipped with a more mechanistic understanding of decisions processes and various neuroscientific tools to measure affective responses (skin conductance responses, pupil dilatation), attentional effects (eye movements, mouse movements), and brain responses elicited by products.

One particularly problematic ethical concern that derives from those new approaches is the ability to target specific individuals or groups of individuals (Stanton et al., 2017) via the systematic monitoring of consumers' behavior, both online and in shops and the use of big data techniques to profile them (Aguirre et al., 2015; Boerman et al., 2017). The goal is to identify the putative needs of categories of consumers in order to focus the marketing strategy on selected goods susceptible to fill those needs. There are several risks associated with this practice, one being an increased consumerism and increased prices paid by consumers (Stanton et al., 2017). Another risk is to exploit the vulnerabilities of individuals. For instance, individuals, with compulsive buying disorders (Black, 2007) are particularly sensitive to encouragements to buy on the web (Rose and Dhandayudham, 2014). Marketing techniques can potentially have detrimental consequences on several groups of the population.

INTER-INDIVIDUAL DIFFERENCES IN DECISION-MAKING AND VULNERABILITY TO MARKETING

Large inter-individual differences exist, both in decision mechanisms and their susceptibility to external influence. During development and aging, individuals tend to make less advantageous choices and are more susceptible to the influence of marketing techniques. Addiction and eating disorders can deeply tamper with the ability of making healthy choices. Recent advances in cognitive psychology and neuroscience can help understand why many individuals struggle in making sound choices.

Children and Adolescents

Compared to adults, adolescents engage more in risky behavior (Steinberg, 2008) and display heightened peer-influence in their daily choices (van Hoorn et al., 2016). The uneven neurodevelopmental trajectories of the brain systems implicated in processing rewards on one side, and those involved in cognitive control on the other can explain these behavioral characteristics (Casey et al., 2008). The hyper-reactivity of the reward system, especially in the striatum is associated with emotional hypersensitivity to rewarding stimuli, faces and socio-emotional stimuli (Galvan et al., 2006; Casey et al., 2008; Hare et al., 2008). By contrast, the maturation of the prefrontal cortex, involved in cognitive control, still continues until about the age of 20 (Gogtay et al., 2004; Shaw et al., 2008).

Younger consumers constitute a substantial part of the market and marketers and advertisers have developed a large spectrum of strategies to reach them (Valkenburg and Cantor, 2001). The interest for marketing in children and adolescents lays in the realization that, in the last decades, they have acquired higher financial independence and more influence in household purchasing decisions. Children develop brand loyalty at an early age (Haryanto et al., 2016), which persists until adulthood. Detrimental effects of advertising on the development of children's consumption habits is well documented (Wilcox et al., 2004). Television commercials targeted at children, in particular, are highly effective (Atkin, 1978; Gorn and Goldberg, 1982). They have been reported to induce unhealthy eating habits, to cultivate a materialistic value system and to be a source of conflicts between children and their parents (Goldberg and Gorn, 1978; Gorn and Goldberg, 1982; Story and French, 2004).

Older Adults

Aging individuals constitute a particularly vulnerable population as well. Older individuals make more disadvantageous decisions, especially in uncertain or changing environments. One exception is the ability to make more farsighted decisions with age (Samanez-Larkin and Knutson, 2015) which can potentially lead to better consumer choices (Zauberman and Urminsky, 2016). However, older adults borrow at higher interest rates and pay more fees to financial institutions than their younger counterparts (Agarwal et al., 2007); they are less consistent in health-related decisions (Löckenhoff and Carstensen, 2007). Most importantly they are more sensitive to deceptive advertising than their younger counterparts (Denburg et al., 2007). Older adults' heightened susceptibility to misleading advertising techniques can be explained by a reduced ability to discriminate between potentially misleading and more truthful advertising claims (Gaeth and Heath, 1987). They tend as well to give higher credit to claims that are repeated. Strikingly, even if they are informed that a claim is false, they will remember it as true a few days later (Skurnik et al., 2005). Decision deficits that arise

with age in variable or uncertain environments might be due to cognitive limitations (Henninger et al., 2010; van de Vijver et al., 2015). Deficits in valuation processes have been also reported at the neural level, as structural changes in frontostriatal pathways are linked to disadvantageous decisions (Samanez-Larkin and Knutson, 2015; van de Vijver et al., 2016).

Inter-Individual Differences in Self-Control

Individuals differ widely in their ability to implement self-control in their daily choices and maintain goal-directed behavior. Economists explain these disparities by considering inter-individual differences in discounting the long term consequences of choice options in the computation of their value (Laibson, 1997; O'Donoghue and Rabin, 1999). Psychologists approach this question by considering the relative difficulty and reliability of representing immediate pleasurable attributes and more abstract and temporally distant attributes of options (Liberman and Trope, 2008). When applied to self-control in dietary choices, eating a chocolate cake rather than an apple can be explained by the overweighing of taste compared to health information. A computational approach showed that up to 39% of the variability in dietary self-control failures can be explained by the speed with which the decision-making circuitry processes basic attributes like taste, versus more abstract attributes such as health (Sullivan et al., 2015). The biological plausibility of this model was supported by the finding that variability in diet success is linked to the relative representation of taste and health attributes in the ventromedial prefrontal cortex (Hare et al., 2009). According to the authors, "these findings provide a rationale for regulating marketing practices that increase the relative ease with which abstract attributes such as health are processed. For example, prominently displaying health information such as calorie counts may allow more rapid integration of health attributes" (Sullivan et al., 2015, p. 133).

In sum, the brain structures involved in motivation and decision-making are the latest to be fully functional during development and decline relatively early with age (Somerville and Casey, 2010; Samanez-Larkin and Knutson, 2015). As a result, maintaining goal-directed behavior in the long term and resisting temptations can be difficult at young age. Later in life, flexibly adapting to changing decision environments can become challenging (Eppinger et al., 2011). During adult life, unhealthy habits can readily form and several biological or societal factors can dysregulate the balance of the decision-making and motivation brain circuitry. Thus, large portion of the population is susceptible to be negatively impacted by marketing techniques and make disadvantageous decision or forming unhealthy habits, at least during certain period of their lives.

ADVERTISING REGULATION

The realization of the increasing potential of neuroscientific knowledge applied to marketing raises a few questions. Does this always represent an advantage to us as a society and as individuals? If not, should (more) regulations be put in place to avert potential damage?

Why Regulate Advertising?

In a world full of temptations carried by pervasive marketing messages, making decisions consistent with one's own goals and preferences requires constant self-control. Extensive research has revealed that self-control often fails when individuals experience emotional distress (Baumeister et al., 1994). Excessive exposure to social norms brought by advertisement can induce emotional distress in vulnerable populations such as addicts or individuals with eating disorders. For instance, exposure to thin models in advertisement induces body-focused anxiety among women (Halliwell and Dittmar, 2004).

Research on the psychological consequences of poverty indicates a link between low income, stress and short-sighted, disadvantageous economical decisions (Haushofer and Fehr, 2014). In addition, financial scarcity causes a reduction in cognitive control (Mani et al., 2013), as well as changes in attention allocation; salient information relative to short-term decisions receive more attention than information concerning the future, which can cause bad economic decisions such as over-borrowing (Shah et al., 2012). Consequently, we might reasonably expect that poorer individuals can be negatively affected by advertising. While positive nudging can elicit people to save more (Karlan et al., 2016), tempting advertising or branding effects can easily lead to over-spending. Whether overexposure to marketing messages is linked to decreased well-being and increased level of stress or emotional distress in the general population is unknown, although some authors suggest it is likely to be the case (Baumeister, 2002; Sullivan et al., 2015). Research investigating this question is crucially needed in order to have a sound scientific dialogue about the "dark side of consumer neuroscience" (Kenning and Plassmann, 2008).

Internet advertising, in particular, potentially constitutes a serious concern. Internet ads are present in the visual field of consumers even when not directly attended. Several studies have shown that the value associated with specific stimuli are retrieved and updated by our reward system even when passively viewed (Lebreton et al., 2009; Tusche et al., 2010; Smith et al., 2014). Passive viewing of products of a specific brand have direct effect on purchase decisions (Ferraro et al., 2009). Additionally, with the generalization of online shopping, ads are present in the visual field of the buyer right at the moment of purchasing decisions. The use of internet data enables the tailoring of adverts by proposing to specific consumers those products they would be more likely to purchase. Online targeted advertising, through the monitoring of people's online behavior triggers an increase in the rate of clicking on the ads as well as higher likelihood of purchase (Boerman et al., 2017), although the size of reported effects varies deeply between academic studies and claims made by advertising agencies.

How to Regulate Advertising?

An efficient and self-regulated market rests on the ability for firms to inform consumers about their novel products and stimulate them to buy those products. Yet, this should not be done at

the expense of individuals' mental, physical or financial health. Neither should marketing strategies drive consumers away from their explicit goals and intentions, such as staying on a diet or reducing their use of products with high environmental impact. While people with strong initial preferences are less likely to see their choice behavior dramatically influenced by marketing techniques, the latter are more efficient on individuals whose preferences have not yet formed such as children, vulnerable groups or individuals with conflicting motivations.

We believe that expanding our knowledge about decision mechanisms and how to modulate them is not inherently problematic as many beneficial applications, for individuals and for the society, can arise. The rehabilitation of addictive disorders is one important application. Nudging, which can be considered as the 'good' counterpart to marketing, relies on very similar theories and techniques to influence individuals' behavior to make it more in line with their intentions. One previously mentioned example is the use of reminders to save money. Another example is the so called 'green-nudging' (Schubert, 2017; Bonini et al., 2018) which prompts people to make ecologically responsible decisions. The key difference between marketing and nudging lies in the very idea of adequacy between the declared intentions of the customer (e.g., follow a specific diet, make ecologically responsible purchases) and the type of manipulation being exerted on their behavior. In addition, nudging is usually initiated by public institutions with the end goal of benefiting the society. For instance, nudging might encourage more ecologically responsible consumption by displaying the environment impact of products, but it will never orient consumers toward a specific brand. Public acceptability of nudging is generally positive (Reynolds et al., 2019) while advertising made by companies motivated by profit is controversial. Therefore, the very idea of transparency from the part of the advertising company and consent from the customer seems crucial. Policy makers could consider empowering citizens by letting them decide whether they accept to be exposed to different types of advertising.

Strikingly, the legal system of several countries has adjudicated that promoting products which threaten public health should be prohibited. Advertisement of products containing tobacco or alcohol is strictly forbidden in many countries. In addition, the branding effect of cigarettes is reduced by including pictures of dramatic health consequences of smoking on packaging. Similarly, attempts to reduce the prevalence of obesity, diabetes and hypertension have been made by trying to limit the effectiveness of advertisements on high caloric food and beverages with associated warning messages. For instance, in 2007 in France, a law was adopted listing categories of nutritive products (e.g., sweets and sodas) whose advertisement had to contain a message suggesting to eat more fruits and vegetable, increase physical activity and reduce salt and sugar intake. Thus, the approach adopted so far to protect the population from potential detrimental effects of advertising focuses on specific products and age groups (mainly children). Nonetheless, as discussed earlier the potential damage of advertising extent to many groups of individuals.

A possibly efficient approach could be to limit the intrusive aspects of the advertising means, in order to allow vulnerable individuals, especially those with compulsive or addictive tendencies, to maintain self-protective strategies. Measures should be taken to prevent advertisement to be forced into the peripheral visual field of individuals attending a nearby focal point of interest. In order to avoid passive viewing, it could entail the prohibition of advertising messages in confined public spaces (e.g., bus stops) and in locations surrounding informative or salient focal point (e.g., information panels). One particularly striking example is the advertisement low-cost airplane companies place on the seat in front of their clients to incite them to buy snacks. Such practice is extremely intrusive as people cannot easily look away. Similarly, if advertisement in magazines would be on their own separate page, rather than next to an informative article, consumers would still have the opportunity of being informed of new products while controlling the degree of exposure to advertisement they are willing to accept. Internet ads could be forced in their own browser tab instead of being placed next to the focus of attention of users. A mandatory op-out option for specific categories of products would also be desirable to help individuals struggling with addictive behavior or eating disorders. The important aspect in this proposition is to allow consumers to regain control in their exposure to advertisement by having them consent to viewing ads through a motor action (such as clicking on the ads tab), rather than forcing passive viewing.

CONCLUSION

Due to our increasing knowledge of decision mechanisms and the increasing efficiency and outreach of communication means, marketing techniques are becoming both intrusive and powerful. The brain circuitry for decision and motivation changes during the lifespan or due to a diversity of contingent and individual factors. Because of our growing understanding of vulnerabilities to external influences, it is perhaps time to address the issue of intrusiveness of advertisement at a societal level and consider regulatory intervention.

AUTHOR CONTRIBUTIONS

NB and ER prepared and validated the manuscript.

FUNDING

This work was funded by the European Research Council (ERC Consolidator Grant 617629).

REFERENCES

Agarwal, S., Driscoll, J. C., Gabaix, X., and Laibson, D. (2007). *The Age of Reason: Financial Decisions Over the Lifecycle*. Cambridge, MA: National Bureau of Economic Research, doi: 10.3386/w13191

Aguirre, E., Mahr, D., Grewal, D., de Ruyter, K., and Wetzels, M. (2015). Unraveling the personalization paradox: the effect of information collection and trust-building strategies on online advertisement effectiveness. *J. Retail.* 91, 34–49. doi: 10.1016/j.jretai.2014.09.005

Armel, K. C., Beaumel, A., and Rangel, A. (2008). Biasing simple choices by manipulating relative visual attention. *Judg. Decision Making* 3, 396–403.

Atkin, C. K. (1978). Observation of parent-child interaction in supermarket decision-making. *J. Mark.* 42, 41–45. doi: 10.1177/002224297804200406

Baumeister, R. F. (2002). Yielding to temptation: self-control failure, impulsive purchasing, and consumer behavior. *J. Consum. Res.* 28, 670–676. doi: 10.1086/338209

Baumeister, R. F., Heatherton, T. F., and Tice, D. M. (1994). *Losing Control: How and Why People Fail at Self-Regulation*, 1 Edn. San Diego: Academic Press.

Black, D. W. (2007). Compulsive buying disorder: a review of the evidence. *CNS Spectr.* 12, 124–132. doi: 10.1017/S1092852900020630

Boerman, S. C., Kruikemeier, S., and Borgesius, F. J. Z. (2017). Online behavioral advertising: a literature review and research Agenda. *J. Advert.* 46, 363–376. doi: 10.1080/00913367.2017.1339368

Bonini, N., Hadjichristidis, C., and Graffeo, M. (2018). Green nudging. *Acta Psychol. Sin.* 50, 814–826. doi: 10.3724/SP.J.1041.2018.00814

Calluso, C., Committeri, G., Pezzulo, G., Lepora, N., and Tosoni, A. (2015). Analysis of hand kinematics reveals inter-individual differences in intertemporal decision dynamics. *Exp. Brain Res.* 233, 3597–3611. doi: 10.1007/s00221-015-4427-1

Casey, B. J., Jones, R. M., and Hare, T. A. (2008). The adolescent brain. *Ann. N. Y. Acad. Sci.* 1124, 111–126. doi: 10.1196/annals.1440.010

Denburg, N. L., Cole, C. A., Hernandez, M., Yamada, T. H., Tranel, D., Bechara, A., et al. (2007). The orbitofrontal cortex, real-world decision making, and normal aging. *Ann. N. Y. Acad. Sci.* 1121, 480–498. doi: 10.1196/annals.1401.031

Eppinger, B., Hämmerer, D., and Li, S.-C. (2011). Neuromodulation of reward-based learning and decision making in human aging. *Ann. N. Y. Acad. Sci.* 1235, 1–17. doi: 10.1111/j.1749-6632.2011.06230.x

Ferraro, R., Bettman, J. R., and Chartrand, T. L. (2009). The power of strangers: the effect of incidental consumer brand encounters on brand choice. *J. Consum. Res.* 35, 729–741. doi: 10.1086/592944

Gaeth, G. J., and Heath, T. B. (1987). The cognitive processing of misleading advertising in young and old adults: assessment and training. *J. Consum. Res.* 14, 43–54. doi: 10.1086/209091

Galvan, A., Hare, T. A., Parra, C. E., Penn, J., Voss, H., Glover, G., et al. (2006). Earlier development of the accumbens relative to orbitofrontal cortex might underlie risk-taking behavior in adolescents. *J. Neurosci.* 26, 6885–6892. doi: 10.1523/JNEUROSCI.1062-06.2006

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., et al. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *PNAS* 101, 8174–8179. doi: 10.1073/pnas.0402680101

Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.* 30, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Goldberg, M. E., and Gorn, G. J. (1978). Some unintended consequences of TV advertising to children. *J. Consum. Res.* 5, 22–29. doi: 10.1086/208710

Gorn, G. J., and Goldberg, M. E. (1982). Behavioral evidence of the effects of televised food messages on children. *J. Consum. Res.* 9, 200–205. doi: 10.1086/208913

Halliwell, E., and Dittmar, H. (2004). Does size matter? the impact of model's body size on women's body-focused anxiety and advertising effectiveness. *J. Soc. Clin. Psychol.* 23, 104–122. doi: 10.1521/jscp.23.1.104.26989

Hare, T. A., Camerer, C. F., and Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* 324, 646–648. doi: 10.1126/science.1168450

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., and Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J. Neurosci.* 28, 5623–5630. doi: 10.1523/jneurosci.1309-08.2008

Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P., and Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. *Proc. Natl. Acad. Sci.* 108, 18120–18125. doi: 10.1073/pnas.1109322108

Haryanto, J. O., Moutinho, L., and Coelho, A. (2016). Is brand loyalty really present in the children's market? A comparative study from Indonesia, Portugal, and Brazil. *J. Bus. Res.* 69, 4020–4032. doi: 10.1016/j.jbusres.2016.06.013

Haushofer, J., and Fehr, E. (2014). On the psychology of poverty. *Science* 344, 862–867. doi: 10.1126/science.1232491

Henninger, D. E., Madden, D. J., and Huettel, S. A. (2010). Processing speed and memory mediate age-related differences in decision making., processing speed and memory mediate age-related differences in decision making. *Psychol. Aging* 25, 262–270. doi: 10.1037/a 0019096

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203. doi: 10.1038/35058500

Karlan, D., McConnell, M., Mullainathan, S., and Zinman, J. (2016). Getting to the top of mind: how reminders increase saving. *Manag. Sci.* 62, 3393–3411. doi: 10.1287/mnsc.2015.2296

Karmarkar, U. R., and Plassmann, H. (2019). Consumer neuroscience: past. present, and future. *Org. Res. Methods* 22, 174–195. doi: 10.1177/1094428117730598

Kenning, P. H., and Plassmann, H. (2008). How neuroscience can inform consumer research. *IEEE Trans. Neural Syst. Rehabil. Eng.* 16, 532–538. doi: 10.1109/TNSRE.2008.2009788

Krajbich, I., Armel, C., and Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* 13, 1292–1298. doi: 10.1038/nn.2635

Krajbich, I., Lu, D., Camerer, C., and Rangel, A. (2012). The attentional drift-diffusion model extends to simple purchasing decisions. *Front. Psychol.* 3:193. doi: 10.3389/fpsyg.2012.00193

Krajbich, I., and Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13852–13857. doi: 10.1073/pnas.1101328108

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *Q. J. Econ.* 112, 443–478. doi: 10.1162/003355397555253

Lebreton, M., Jorge, S., Michel, V., Thirion, B., and Pessiglione, M. (2009). An automatic valuation system in the human brain: evidence from functional neuroimaging. *Neuron* 64, 431–439. doi: 10.1016/j.neuron.2009.09.040

Liberman, N., and Trope, Y. (2008). The psychology of transcending the here and now. *Science* 322, 1201–1205. doi: 10.1126/science.1161958

Lim, S.-L., O'Doherty, J. P., and Rangel, A. (2011). The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. *J. Neurosci.* 31, 13214–13223. doi: 10.1523/JNEUROSCI.1246-11.2011

Löckenhoff, C. E., and Carstensen, L. L. (2007). Aging, emotion, and health-related decision strategies: motivational manipulations can reduce age differences. *Psychol. Aging* 22, 134–146. doi: 10.1037/0882-7974.22.1.134

Mani, A., Mullainathan, S., Shafir, E., and Zhao, J. (2013). Poverty impedes cognitive function. *Science* 341, 976–980. doi: 10.1126/science.1238041

Milosavljevic, M., Navalpakkam, V., Koch, C., and Rangel, A. (2012). Relative visual saliency differences induce sizable bias in consumer choice. *J. Consum. Psychol.* 22, 67–74. doi: 10.1016/j.jcps.2011.10.002

Navalpakkam, V., Koch, C., Rangel, A., and Perona, P. (2010). Optimal reward harvesting in complex perceptual environments. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5232–5237. doi: 10.1073/pnas.0911972107

O'Donoghue, T., and Rabin, M. (1999). Doing it now or later. *Am. Econ. Rev.* 89, 103–124. doi: 10.1257/aer.89.1.103

Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., and Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *PNAS* 112, 4170–4175. doi: 10.1073/pnas.1415250112

Pieters, R., and Wedel, M. (2004). Attention capture and transfer in advertising: brand, pictorial, and text-size effects. *J. Mark.* 68, 36–50. doi: 10.1509/jmkg.68.2.36.27794

Plassmann, H., Ambler, T., Braeutigam, S., and Kenning, P. (2007). What can advertisers learn from neuroscience? *Int. J. Advert.* 26, 151–175. doi: 10.1080/10803548.2007.11073005

Plassmann, H., and Karmarkar, U. R. (2015). "Consumer neuroscience: revealing meaningful relationships between brain and consumer behavior," in *The Cambridge Handbook of Consumer Psychology*, eds M. I. Norton, D. D. Rucker, and C. Lamberton, (Cambridge, MA: Cambridge University Press), 152–179. doi: 10.1017/CBO9781107706552.006

Rangel, A., and Clithero, J. A. (2014). "The computation of stimulus values in simple choice," in *Neuroeconomics*, ed. P. W. Glimcher, (Amsterdam: Elsevier), 125–148. doi: 10.1016/B978-0-12-416008-8.00008-5

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev.* 85, 59–108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion decision model: current issues and history. *Trends Cogn. Sci.* 20, 260–281. doi: 10.1016/j.tics.2016.01.007

Reutskaja, E., Nagel, R., Camerer, C. F., and Rangel, A. (2011). Search dynamics in consumer choice under time pressure: an eye-tracking study. *Am. Econ. Rev.* 101, 900–926. doi: 10.1257/aer.101.2.900

Reynolds, J. P., Archer, S., Pilling, M., Kenny, M., Hollands, G. J., and Marteau, T. M. (2019). Public acceptability of nudging and taxing to reduce consumption of alcohol, tobacco, and food: a population-based survey experiment. *Soc. Sci. Med.* 236, 112395. doi: 10.1016/j.socscimed.2019.112395

Roitman, J. D., and Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* 22, 9475–9489. doi: 10.1523/JNEUROSCI.22-21-09475.2002

Rose, S., and Dhandayudham, A. (2014). Towards an understanding of Internet-based problem shopping behaviour: the concept of online shopping addiction and its proposed predictors. *J. Behav. Addict.* 3, 83–89. doi: 10.1556/JBA.3.2014.003

Ruff, C. C., and Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nat. Rev. Neurosci.* 15, 549–562. doi: 10.1038/nrn3776

Samanez-Larkin, G. R., and Knutson, B. (2015). Decision making in the ageing brain: changes in affective and motivational circuits. *Nat. Rev. Neurosci.* 16, 278–289. doi: 10.1038/nrn3917

Schubert, C. (2017). Green nudges: do they work? Are they ethical? *Ecol. Econ.* 132, 329–342. doi: 10.1016/j.ecolecon.2016.11.009

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27. doi: 10.1152/jn.1998.80.1.1

Shah, A. K., Mullainathan, S., and Shafir, E. (2012). Some consequences of having too little. *Science* 338, 682–685. doi: 10.1126/science.1222426

Shaw, P., Kabani, N. J., Lerch, J. P., Eckstrand, K., Lenroot, R., Gogtay, N., et al. (2008). Neurodevelopmental trajectories of the human cerebral cortex. *J. Neurosci.* 28, 3586–3594. doi: 10.1523/JNEUROSCI.5309-07.2008

Shimojo, S., Simion, C., Shimojo, E., and Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nat. Neurosci.* 6, 1317–1322. doi: 10.1038/nn1150

Skurnik, I., Yoon, C., Park, D. C., and Schwarz, N. (2005). How warnings about false claims become recommendations. *J. Consum. Res.* 31, 713–724. doi: 10.1086/426605

Smith, A., Bernheim, B. D., Camerer, C., and Rangel, A. (2014). Neural activity reveals preferences without choices. *Am. Econ. J. Microecon.* 6, 1–36. doi: 10.1257/mic.6.2.1

Somerville, L. H., and Casey, B. (2010). Developmental neurobiology of cognitive control and motivational systems. *Curr. Opin. Neurobiol.* 20, 236–241. doi: 10.1016/j.conb.2010.01.006

Stanton, S. J., Sinnott-Armstrong, W., and Huettel, S. A. (2017). Neuromarketing: ethical implications of its use and potential misuse. *J. Bus. Ethics* 144, 799–811. doi: 10.1007/s10551-016-3059-0

Steinberg, L. (2008). A social neuroscience perspective on adolescent risk-taking. *Curr. Direct. Risk Decis. Making* 28, 78–106. doi: 10.1016/j.dr.2007.08.002

Story, M., and French, S. (2004). Food advertising and marketing directed at children and adolescents in the US. *Int. J. Behav. Nutr. Phys. Act.* 1:3. doi: 10.1186/1479-5868-1-3

Sullivan, N., Hutcherson, C., Harris, A., and Rangel, A. (2015). Dietary self-control is related to the speed with which attributes of healthfulness and tastiness are processed. *Psychol. Sci.* 26, 122–134. doi: 10.1177/0956797614559543

Summerfield, C., and Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: neural and computational mechanisms. *Front. Neurosci.* 6:70. doi: 10.3389/fnins.2012.00070

Tobler, P. N., Fiorillo, C. D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science* 307, 1642–1645. doi: 10.1126/science.1105370

Towal, R. B., Mormann, M., and Koch, C. (2013). Simultaneous modeling of visual saliency and value computation improves predictions of economic choice. *Proc. Natl. Acad. Sci.* 110, E3858–E3867. doi: 10.1073/pnas.1304429110

Tusche, A., Bode, S., and Haynes, J.-D. (2010). Neural responses to unattended products predict later consumer choices. *J. Neurosci.* 30, 8024–8031. doi: 10.1523/JNEUROSCI.0064-10.2010

Valkenburg, P. M., and Cantor, J. (2001). The development of a child into a consumer. *J. Appl. Dev. Psychol.* 22, 61–72. doi: 10.1016/S0193-3973(00)00066-6

van de Vijver, I., Ridderinkhof, K. R., and de Wit, S. (2015). Age-related changes in deterministic learning from positive versus negative performance feedback. *Aging Neuropsychol. Cogn.* 22, 595–619. doi: 10.1080/13825585.2015.1020917

van de Vijver, I., Ridderinkhof, K. R., Harsay, H., Reneman, L., Cavanagh, J. F., Buitenweg, J. I. V., et al. (2016). Frontostriatal anatomical connections predict age- and difficulty-related differences in reinforcement learning. *Neurobiol. Aging* 46, 1–12. doi: 10.1016/j.neurobiolaging.2016.06.002

van Hoorn, J., Fuligni, A. J., Crone, E. A., and Galván, A. (2016). Peer influence effects on risk-taking and prosocial decision-making in adolescence: insights from neuroimaging studies. *Curr. Opin. Behav. Sci.* 10, 59–64. doi: 10.1016/j.cobeha.2016.05.007

Wilcox, B. L., Kundel, D., Cantor, J., Dowrick, P., Linn, S., and Palmer, E. (2004). *Report of the APA Task Force on Advertising and Children*. Washington, D.C: American Psychological Association, doi: 10.1037/e539692009-001

Zauberman, G., and Urminsky, O. (2016). Consumer intertemporal preferences. *Curr. Opin. Psychol.* 10, 136–141. doi: 10.1016/j.copsyc.2016.01.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership