

BALKAN JOURNAL OF PHILOSOPHY

Vol. 10, Issue 1, 2018

TABLE OF CONTENTS

Special issue devoted to the topic of “Neurolaw: its current state and perspectives for future development”

Note from the Editorial Board / 3

Articles

Invited paper

Intellectual Disability, Brain Damage, and Group-to-Individual Inferences:
How the U.S. Court System Uses Neuroscience Data / 5

Valerie Gray Hardcastle

Invited paper

Looping effects of neurolaw, and the precarious marriage between neuroscience and the
law / 17

Toma Strle, Olga Markič

Psychiatry and neurolaw

An Essay on the Mind-Brain Problem and Legal Proof / 27

Drozdstoy St. Stoyanov

Naturalizing the subjective side of the crime: a few introductory remarks on the role of
consciousness in criminal law based on American and Polish examples / 37

Bartosz Janik, Maciej Próchnicki

Applying Confirmation Theory to the Case against Neurolaw / 45

Anton Donchev

Freedom, Responsibility and Jurisprudence / 55
Hari Narayanan V

Retributivist Theory of Punishment: Some Comments / 63
Adebayo Aina

SPECIAL ISSUE OF “BALKAN JOURNAL OF PHILOSOPHY”

NEUROLAW: ITS CURRENT STATE AND PERSPECTIVES FOR FUTURE DEVELOPMENT

The term “neurolaw” was introduced in the 1990s (Taylor, Harp and Elliott, 1991) to name the emerging interdisciplinary area of research aimed at exploring the possible effects of achievements in neuroscience on the traditional legal systems and legal practices of democratic countries. Neurolaw was born in the US, but has been slowly spreading throughout the rest of the world (Spranger, 2012).

Since its inception, the research area of neurolaw has spawned numerous controversies which still divide philosophers, neuroscientists, and legal scholars. One of the most heated debates is centered around the question “Should we change current legal systems in order to make them compatible with the scientific view of human nature provided by contemporary neuroscience?” On one side of this debate are those who believe that the existing legal systems are based on outdated concepts which must undergo significant changes, prompted by the latest findings in neurobiology (viz. David Eagleman). On the other side are critics of neurolaw who insist that the legal systems should not be transformed under the pressure of neuroscientific discoveries, insofar as rules and standards in law are essentially dependent on descriptions and explanations of behavior in folk-psychological terms (viz. Stephen Morse).

The central debate about the relevance of neuroscience to law has raised additional issues which many think are of no less importance. On the one hand, there are practical questions about the best way to prepare the neuroscientist for the courtroom (Jones, et al. 2013), or about the type

of neuroscientific evidence that may be admitted in court (e.g. Keckler, 2005). On the other hand, there are philosophical questions concerning the impact of neuroscientific findings on the problem of free will and the related notion of legal responsibility (Greene and Cohen, 2004), and on the mind-brain problem and its relevance to the question of whether law should stick to folk-psychological accounts of behavior (Glan-non, 2009).

The aim of this special issue of the *Balkan Journal of Philosophy* is to investigate the current state of debates on neurolaw, and to assess plausible perspectives for the future development of this new and socially significant interdisciplinary area of research. Therefore, we call for papers addressing, but not limited to, the following topics:

- The current state of neurolaw and the debates on it in different countries;
- New and promising lines of research in the field of neurolaw;
- The roles of different disciplines in neurolaw (viz. psychology, psychiatry, computer science, social science, philosophy, etc.);
- The main neuroscientific challenges to traditional legal systems and practices;
- Ways in which current legal systems could benefit from neuroscience;
- Types of neuroscientific evidence admissible in court;
- Whether law essentially depends on folk-psychological accounts of behavior;

- Neurolaw, free will, and legal responsibility;
- Neurolaw and the mind-brain problem.

References

Glannon W (2009) Our brains are not us. *Bioethics*, 23(6): 321–329.

Greene J, Cohen J (2004) For the Law, Neuroscience Changes Nothing and Everything. *Philosophical Transactions: Biological Sciences*. 359(1451): 1775–1785.

Jones O, Wagner A, Faigman D, Raichle M (2013) Neuroscientists in court. *Nature Reviews Neuroscience*, 14(10): 730–736.

Keckler C N (2005) Cross-examining the Brain: A Legal Analysis of Neural Imaging for Credibility Impeachment. *Hastings LJ*, 57: 509–556.

Spranger T (ed.) (2012) *International Neurolaw: A Comparative Analysis*. Berlin: Springer.

Taylor J, Harp J, Elliott T (1991) Neuropsychologists and Neurolawyers. *Neuropsychology*, 5(4): 293–305.

From the Editorial Board of
Balkan Journal of Philosophy

Articles

Invited paper

INTELLECTUAL DISABILITY, BRAIN DAMAGE, AND GROUP-TO-INDIVIDUAL INFERENCES: HOW THE U.S. COURT SYSTEM USES NEUROSCIENCE DATA

Valerie Gray Hardcastle

Professor of Philosophy, Psychology, and Psychiatry & Behavioral Neuroscience
Scholar-in-Residence, Weaver Institute for Law and Psychiatry
University of Cincinnati
e-mail: hardcave@ucmail.uc.edu

Abstract:

In this essay, I home in on the difficulties with group-to-individual (G2i) inferences in neuroscience and how they impact the legal system. I briefly outline how cognitive shortcutting can distort legal decisions, and then turn my attention to G2i inferences, with a special focus on issues of intellectual disability and brain damage. I argue that judges and juries are not situated to appreciate the nuances in brain data and that they are required to make clinical decisions without clinical training. As a result, they effectively ignore those responsibilities and simply decide cases in virtue of what they already believe to be true. How judges actually make decisions in high-stakes criminal cases is troubling, but they are also hamstrung in a variety of ways.

Key words: neuroscience, criminal cases, intellectual disability, brain damage.

neuroscience evidence in criminal cases is the focus of this essay.

In particular, I shall home in on the difficulties with group-to-individual (G2i) inferences in neuroscience and how they impact the legal system. Below, I first briefly outline how cognitive shortcutting can distort legal decisions, and then turn my attention to G2i inferences, with a special focus on issues of intellectual disability and brain damage. I argue that judges and juries are not situated to appreciate the nuances in brain data and that they are required to make clinical decisions without clinical training. As a result, they effectively ignore those responsibilities and simply decide cases in virtue of what they already believe to be true. How judges actually make decisions in high-stakes criminal cases is troubling, but they are also hamstrung in a variety of ways.

1. Introduction

Judicial systems operate under the presumption that judges and juries are fully rational, that they simply apply the rules of the case to the facts at hand to arrive at a just verdict. But, of course, judges and juries are also people. And, just like the rest of us, they all have psychological foibles, weaknesses, prejudices, and biases (see also Berman and Hafner 1989, Bruner 2003, Hardcastle 2017, Hardcastle, Kitzmiller, and Lahey 2018, McCarty 1997, Susskind 1986). How these imperfections interact with

Bias in Decision-Making

We all use cognitive shortcuts to help speed and streamline our decision-making process. In particular, people actively seek information from their environments to confirm what they already take to be true and actively discount or ignore information that disconfirms their intuitions. For example, if judges or jurors already assume that mentally ill defendants are, say, not rule-abiding, then they will preferentially, albeit unconsciously, seek evidence in the trial to confirm their suspicion. They will interpret any ambiguous data in such a way that aligns with their

implicit belief that the mentally ill have tendencies toward violence.

Recent work in social psychology indicates that how individuals assign responsibility is correlated with prior judgments of what counts as being morally bad, which are in turn dependent upon other, larger, social and cultural factors. While there is a long and deep history of this research in social psychology, Mark Alicke's culpable control model of blame presents a good exemplar of this approach (Alicke 2000, 2008; Alicke et al. 2008a; 2008b). "Culpable control" refers to the fact that our desire to blame someone intrudes on our assessments of that person's ability to control his or her thoughts or behavior. Deciding that someone is responsible for an act, which is taken to be the conclusion of a judgment, is actually part of our psychological process of assessing blame. Starting with a spontaneous negative reaction can lead to hypothesizing that the source of an action is blameworthy as well as to an active desire to blame that source (just as starting with a spontaneous positive reaction can lead to looking for exculpatory reasons). These spontaneous reactions, in turn, skew interpretations of the available evidence such that it supports a blame hypothesis. Individuals highlight evidence that indicates negligence, recklessness, impure motives, or a faulty character, and ignore evidence that suggests otherwise. In other words, instead of dispassionately judging whether someone is responsible, individuals validate their spontaneous reactions of blameworthiness (or innocence).

This, of course, is just another instance of confirmation bias, the tendency to search for, interpret, and remember information in such a way that it confirms one's preconceptions (Watson 1960; see also Kahneman 2011, Tversky and Kahneman 1974). This well-known type of systematic bias arises in many cases of inductive reasoning. Modern interpretations of this phenomenon suggest that it is a result of informal heuristics individuals use to make decisions more efficiently when confronted with a significant amount of information that needs to be processed quickly (Friedrich 1993, Hergovich et al. 2010, Kunda 1999, Maccoun 1998, Matlin 2004, Nickerson 1998, Oswald and Grosjean 2004). Jon Hanson in particular has argued that these sorts of heuristics often impair policy-makers and our judicial system (Benforado and Hanson 2005, 2008a, 2008b, 2008c; Chen and Hanson, 2004a, 2004b; Hanson 2012, Hanson and McCann 2008, Hanson and Yosifin 2003, 2004).

Importantly, recent studies have shown that judges and juries are not only biased in the same ways as the general population in the United States, but that these biases do in fact affect their memories and judgments (Anwar et al. 2012, Eberhardt et al. 2006, Goff et al. 2008, Levinson 2007). Crucially, Uhlmann and Cohen concluded that, "when people feel that they are objective, rational actors, they act on their group-based biases more rather than less" (2007, p. 221). As Kang et al. states: "believing oneself to be objective is a prime threat to objectivity" (2012, p. 184). In addition, those who are cognitively overloaded by complex tasks or time pressures — like judges with overly full dockets — show more bias, likely a byproduct of individuals' reduced level of control over responses due to increasing demands on attentional processes (Payne 2006, Reskin 2005). In short, there is now a considerable body of research that increasingly calls into question whether judges and juries have the ability to render truly objective verdicts (Levinson 2007, Levinson et al. 2010, Levinson et al. 2014, Reynolds 2013).

If a judge or jury is already predisposed to believe that a defendant is guilty, and they selectively interpret evidence to support intuitive judgments of culpability, then they would be more likely to view ambiguous evidence as aggravating. In other words, if everything else is held constant, when presented with ambiguous neuroscience evidence, judges and juries would be more likely to find a defendant they are biased against guilty and would be more likely to sentence that person to a longer term.

Neuroscience Data in the Courtroom

Even though their numbers are still not huge (Shen 2017), neuroscience data have now appeared in enough criminal cases to allow for some general conclusions regarding how judges and juries are using this material in their decisions. It is one thing to believe that a brain defect can lead to violence; it is quite another to excuse a guilty defendant due to a faulty brain. I have systematically examined the conditions under which we appear willing to do so and describe some of my conclusions below.

I culled a database of appellate criminal cases in the United States with judicial decisions released from 1 October 2015 through 30 April 2016 that reference brain data in the decision. Searches on the WestLaw database using the parameters "brain OR neuro!" resulted in 3,213 cataloged appellate decisions released during the study timeframe. Decisions that referenced only brain data from victims were

eliminated, as were *pro forma* decisions that reduced or remanded sentences for minors, following recent Supreme Court rulings on these matters. That left 1,018 appellate decisions that cited brain data in the finding. Using mixed-method quantitative and qualitative analyses, the outcomes of decision were analyzed along 16 dimensions, including proffered brain data, putative use of the evidence, type of crime, sentence, age of defendant, and argument structure.

It should be noted that legal databases only contain a subset of all appellate decisions and, of course, not all judicial decisions are appealed. Moreover, the decisions that get stored in a database are those with large consequences, are especially complex, or have interesting and larger judicial implications. Therefore, the cases I collected necessarily comprise only a portion of the decisions from the relevant year that relied on brain data and are skewed away from smaller, lower impact, decisions. However, because brain data are costly and often difficult to procure, they are only likely to be proffered as evidence in higher impact cases anyway. And higher impact cases are more likely to be appealed. Hence, we have some reasons to accept the conclusions adumbrated below as generalizable to many cases in which brain data are proffered as evidence.

Equivocal Evidence

How to connect brain data to criminal responsibility is a decision for judges and juries, and not for any testifying expert. Therefore, neuroscience data are inherently equivocal in their meaning. In cases in which mitigating evidence could make a difference in outcome, neuroscience data can serve at the juries' or judges' pleasure: it can mitigate when the defendant is seen as a person deserving of mercy, but it can aggravate when the defendant is seen as someone bad.

To illustrate by one example from the database: Katila Ann Jean Nash was found guilty of first-degree murder with the special circumstances of burglary and sentenced to 25 years to life for being present when a murder occurred (*People v. Nash* 2015). She was 15 years old at the time of the offense. Nash, her older sister Angelique Elandra Nash, and her friend David Deshawn Moses went to the home of 81 year-old Dorothy Session, whom they did not know. When Session unexpectedly opened her back door, Moses asked to use her phone, and she willingly let Nash and Moses into her home, while Nash's sister remained in the street. Moses confessed to hitting Session at least twice,

who died later that evening from blunt force trauma. He explained that he had intended to steal money in order to get food and that he had not planned on killing anyone. Nothing, however, was taken from her home. Nash was in the house when Session was being beaten, but she did not strike her. Crying, she ran to find her sister.

At Nash's trial, a forensic psychologist testified that her IQ was 76, which falls in the range "between below average and mental retardation or deficiency." He also diagnosed her with ADHD, anxiety, depression, obsessive-compulsive disorder, PTSD, and "mixed receptive expressive language disorder." Her oral comprehension and her ability to listen and then immediately repeat back a story was at the second grade level; her ability to recall the story the next day was at a first grade level. In addition, he indicated that she has "difficulty processing more than one stimulus at a time," had problems maintaining a sense of self, and that her behaviors, attitudes, and plans were quite erratic. He concluded that she did not "have the reasoning ability or the words needed to understand and form or carry out plans." During a competency hearing, it was also revealed that Nash had been a dependent on the court since she was 10 years old due to parental neglect and abuse. Her mother was an addict and had used methamphetamine when she was pregnant with Nash. One side effect of prenatal exposure to methamphetamine is abnormal brain development. Nash's father had molested her and also abused drugs.

However, in contradistinction to the testimony, the court found that there were "no factors or circumstances in mitigation," and, in addition, the "seriousness and violent nature of the offense" (in which Nash did not participate) made Nash an unsuitable candidate for probation. In its conclusion, the court noted that "what is very striking ... in listening to the evidence of the case that though she was aware what was happening in the home, absolutely no attempt was made to help the defenseless elderly victim. *She still does not understand ... her actions*" (italics added), which of course was exactly the defense's point. The court's interpretation of the proffered neuroscience, psychological, and psychiatric evidence did not diminish its assessment of her culpability, as the defense intended for it to do.

I submit that the description of Nash's life history and her mental deficiencies fits the stereotype of violent female convicts: an addicted mother, an abusive father, a ward of the state, undereducated, low

IQ, defiant attitude, emotional problems, learning disabilities, behavioral difficulties, and African American. Nash embodies the very cultural paradigm for female violent offenders (Barnett 2006, Chesney-Lind 1999, Farr 1997, Huckerby 2003, Sarat 1993, Wilczynski 1991; see also Salisbury and Van Voorhis 2009), and she definitely does not resemble members of a jury.

I suggest that brain data aggravate instead of mitigate when they support judges' and juries' biases regarding who is "bad" and who is not (see also Hardcastle, Kitzmiller, and Lahey 2018). I believe that these tendencies to rely on familiar tropes especially come into play when confronted with defendants who have done truly terrible things, who defy social expectations regarding how people are supposed to behave under duress, and about whom there is little understanding of their actions and motives. It is easier to point to a brain deficit as the culprit than to try to sort out what the implications of having a broken brain might be on their minute-by-minute decisions and actions.

Being a victim of abuse, being intellectually disabled and uneducated, or having a mental illness are all familiar ways of explaining to ourselves why someone might do bad things. At the same time, because all these things can permanently affect the brain, it can be challenging to see the violent behavior as anything other than an expression of a core character. And those with a bad character are perhaps punished more harshly than those who might merely find themselves in unfair situations. This is the idea I explore below.

Group-to-Individual Arguments

Note that the neuroscience evidence referenced in Nash's case was specific to her brain. In other words, in her case, the neuroscientific and psychiatric data were being used diagnostically. Medical doctors and clinical psychologists aim to identify and describe the ailments of their individual patients such that they can treat them; they diagnose. They also can provide testimony about their patients. That is, they use scientific facts and figures to discuss the specific attributes of a particular individual. And courts have long allowed such diagnostic testimony. In the U.S., medical experts testify "to a reasonable degree of medical certainty" (cf., Lewin 1998) that their patients, e.g., were psychotic at the time of the offense and therefore had no appreciation of the wrongfulness of the purported act.

In contrast, research in the brain and behavioral sciences is focused on creating averaged data sets. It is interested in the central tendencies across groups. And courts often attempt to use averaged data that express central tendencies to make individual-level inferences as well (Buckholz and Faigman 2015; see also Faigman et al. 2014). And medical professionals often try to use these group trends to make judgments about individuals, too. However, these group-to-individual (G2i) inferences may not be warranted, for it can be difficult to diagnose a behavioral and cognitive deficiency in a single individual based on group comparisons (see Treadway and Buckholtz 2011).

We see examples of similar awareness of the faultiness of G2i arguments in our database. For example, Dr. Krusz diagnosed defendant Robert Leslie Roberson with a post-concussion syndrome, a type of traumatic brain injury (*Roberson v. Stephens* (2015)). He testified that the syndrome would affect Roberson's impulse control, reasoning capacity, and his ability to act intentionally or knowingly. He also testified that he could relate the diagnosis to the relevant culpable mental states under the law. But the Court of Criminal Appeals in Texas concluded that he did not deliver on that promise, for he could not connect his testimony to "Roberson's ability [in particular] to form knowledge or intent." Thus, they concluded that "poor impulse control [in general] is not relevant as to [Roberson's specific] ability to form the requisite *mens rea* for the offense."

While G2i inferences are not common in criminal cases, they do appear in sufficient numbers to draw at least some tentative conclusions about how they function in them. In our database, we determined that there were 180 G2i decisions, approximately 18% of the total. Overall, 38 out of the 180 G2i cases were decided at least in part in favor of the appellee. These cases either altered the defendant's sentence in some fashion or were remanded back to a lower court for reconsideration. This represents a 22% rate of success for the prisoners. In addition, 25 defendants, out of the 113 convicted on some sort of murder charge (22%) won their appeal, which is a comparable rate of success regarding appeals concerning lesser charges (15 out of 67 cases or a 22% success rate).

It turns out that neither of these statistics aligns with previous analyses of neuroscience-based arguments in criminal cases. Previous investigations into the rate of prisoner success in appellate cases using brain data indicate that it hovers around 10% (Catley

and Claydon 2015, Chandler 2015, de Kogel and Westgeest 2015, Denno 2015, Gaudet and Marchant 2016, Hardcastle 2017). Moreover, they also indicate that the severity of charge is strongly correlated with the increasing failure of neuroscience-based arguments (Denno 2015, Hardcastle 2017, Morse 2012). That is, the more serious the charge, the less likely the neuroscience data would make a difference in outcome. The immediate and obvious question is: Why do we see such different and counter-intuitive results with G2i arguments? Looking more closely at the data helps to answer this question.

The G2i arguments utilized in these cases can be divided into four primary subject categories: Organic Brain Damage or Mental Illness, Substance Abuse, Juveniles, and Intellectual Disabilities. Of the 83 cases in which defendants were categorized with brain damage or mental illness, 11 of them resulted in a decision favorable to the defendant, for a 13% success rate for appellees. Similarly, of the 26 cases of substance use, five cases were decided in favor of the appellee, which is a 19% success rate. However, for the 28 cases that involved juvenile defendants, nine were in favor of the defendant — a 32% success rate. And of the 45 cases in which the defendant was categorized with an intellectual disability, 14 resulted in a change for the prisoner, a success rate of 31%. The difference seems to lie in the particular category to which the prisoner is alleged to belong.

On the surface it appears easy to see why: The U.S. Supreme Court has drawn bright lines around how to treat defendants who are under the age of 18 or intellectually disabled. And these bright lines make it easier for courts to apply group data to individual defendants. Ultimately, however, I shall maintain that these appearances are deceiving in important ways. Here I shall be focusing on cases of intellectual disability and organic brain damage.

Intellectual Disability G2i Arguments

Just as with juveniles tried as adults, the Supreme Court has recently drawn lines around those with intellectual disabilities. In 2002, the United States Supreme Court ruled that executing those with (what was then called) “mental retardation” is unconstitutional, as it violates the Eighth Amendment, which prohibits cruel and unusual punishments (*Atkins v. Virginia*). This decision overturned a previous decision by the Supreme Court in 1989, in which it upheld the constitutionality of executing prisoners with intellectual disabilities (*Penry v.*

Lynaugh). At the time, the Court said that “mental retardation” can be a mitigating factor, but there was no “national consensus” opposing the execution of such prisoners. At the time, only two states (Maryland and Georgia) and the federal government prohibited executing prisoners with intellectual disabilities. However, between 1989 and 2002, 16 more states added such laws (and an additional 19 forbade capital punishment in its entirety). The Court interpreted the Eighth Amendment’s prohibition on cruel and unusual punishment in light of the “evolving standards of decency that mark the progress of a mature society,” and looked to state legislatures for indicators of these standards. Consequently, since “most of the legislatures that have recently addressed the matter” have rejected the death penalty for these offenders, “a national consensus has [now] developed against it.”

The question then becomes how to determine whether someone is intellectually disabled. At first, the Court left that decision to states. But in 2014, the Supreme Court declared that a “rigid rule” for counting as disability, like an IQ score below 70, “disregards established medical practice.” Instead low IQ scores should be combined with additional evidence of adaptive deficits and the presence of both before the age of 18 (*Hall v. Florida* 2014). Most recently, the Court has clarified that states must use standards “informed by the medical community’s diagnostic framework” (*Moore v. Texas* 2017). In particular, because Texas’s standards had relied on stereotypes, lay testimony, and were “untied to any acknowledged source,” they were unconstitutional.

Of course, it is no surprise that there now would be a plethora of appeals based on *Hall*, as prisoners seek to be declared intellectually disabled using standards other than what was used in their original trial in an effort to reduce their sentence. And now we have a good explanation for why some G2i appellate cases are more successful than appellate cases that rely on neuroscience data for diagnosis: The courts used G2i arguments to draw bright lines around particular categories of defendants, and then all the judicial system has to do is determine which defendants fall under those categories.

Of course, the courts are not deciding these cases on the basis of valid G2i arguments; instead, they are relying on precedent and simple categorical assessments. As Scalia noted in his dissent in *Atkins*: the idea that no one who is “mentally retarded can have sufficient moral responsibility to be subjected

to capital punishment for any crime” is an “implausible ... sociological ... conclusion.” And later: “Surely culpability, and deservedness of the most severe retribution, depends not merely (if at all) upon the mental capacity of the criminal (above the level where he is able to distinguish right from wrong) but also upon the depravity of the crime – which is precisely why this sort of question has traditionally been thought answerable not by a categorical rule of the sort the Court today.”

When a bright line is not a bright line. However, to stop here does not capture all the nuances of these cases. In particular, it overlooks the great resistance in the judicial system to embrace the implications of these rulings. Absent a bright line, judges and juries seem to feel free to interpret the relevant science to support their prejudices. And sometimes they do so even with the bright line.

Consider Kenneth Williams of Arkansas, who was executed in April 2017. He had a full-scale IQ score of 70 with a history of severe “learning disabilities” and “neuropsychological problems,” including likely brain damage. He experienced “significant head injuries,” and had “memory problems, problems focusing attention, deficiencies of judgment and reasoning, problems with reading comprehension, problems with comprehending oral instructions and problems with mental flexibility, which is the ability to shift from one task to another.” As one expert witness put it: “His brain is not working the way it should” (*Williams v. State* 2007). And yet, *the same Supreme Court that authored the Atkins decision ultimately denied him an Atkins hearing (Williams v. Kelley* 2017).

Bright lines might be well defined, but even they are not enough to fundamentally alter sentencing decisions in the United States. Despite very clear guidelines to the contrary, courts (even the U.S. Supreme Court) still have a tendency to revert to what they have done in the past.

We find similar patterns throughout our curated G2i cases. For example, Paul Gilbert Devoe was sentenced to death for intentionally shooting and killing two people using a stolen gun (*Devoe v. State* 2011). At trial, a clinical neuropsychologist testified that the “Petitioner functions in many ways like a person with mental retardation in that his intellectual and cognitive abilities are similar to a person with mental retardation and he displays deficits in areas of adaptive functioning, specifically an inability to live independently and a history of poor academic performance.” Further evaluation revealed that

Devoe’s abnormal brain functioning was likely due to “exposure to alcohol in utero (also known as Fetal Alcohol Syndrome and Fetal Alcohol Effects).” In particular, Devoe showed “impulsivity, low verbal skills and antisocial behavior,” as well as “more impairments in verbal skills than nonverbal skills” and some degree of psychoses. The appellate judge agreed that the defendant’s “trial counsel presented a wealth of potentially mitigating evidence during the punishment phase of [his] capital murder trial, including evidence showing [he] suffers from fetal alcohol syndrome/fetal alcohol effects, diminished brain function, psychosis, low intellectual functioning, and a long history of alcohol and drug abuse.” Nonetheless, jury still found death was appropriate sentence.

Double-edged sword and intellectual disabilities. I suggest that, in the cases describe above, the proffered brain data served as a “double-edged sword” in both suggesting that the defendants were not as responsible for their actions as an intellectually normal adult might be, but also giving license to the judge and jury to treat them as incorrigible and dangerous. Deborah Denno (2015, 2017) has argued that neuroscience data is in fact not used 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” — in criminal cases (Denno 2015, p. 503). However, as I discuss below, while we have never found a judge explicitly writing that a prisoner is more likely to be dangerous because of a brain defect, the results of the G2i-based appeals for brain damage (55 of the 180 cases), especially for defendants sentenced to death or life without parole, strongly suggest that the triers of fact actually do believe that these deficits enhance dangerousness.

Indeed, in *Atkins v. Virginia*, the Court specifically spoke to this issue, noting that the intellectually disabled face “a special risk of wrongful execution” in part because “reliance on mental retardation as a mitigating factor can be a two-edged sword that may enhance the likelihood that the aggravating factor of future dangerousness will be found by the jury.” And later in *Cullen v. Pinholster* (2011) the Court reiterated the principle, noting that, “mitigating evidence can be a ‘two-edged sword’ that juries might find to show future dangerousness.”

Certainly, defense attorneys and judges from our set of brain damage G2i decisions believe in the potential for neuroscience data to cut both ways. For example, in 1998, Daniel Anthony Lucas was sen-

tenced to death for the murders of three members of the Moss family during a botched robbery in Jones County, Georgia (*Lucas v. State* 2001). He appealed, arguing that his counsel was ineffective because they failed to fully present mitigating evidence. However, in the end, the Supreme Court of Georgia agreed that it was reasonable for trial counsel to be concerned that this evidence would be a “two-edged sword,” ultimately “undermining the argument that he had endured a tough life but was largely an ‘innocent’ until [others] influenced him” (*Lucas v. Warden* 2014). The court also noted that “the Supreme Court has recognized that evidence of previous abuse and addiction can be mitigating in that it may offset moral culpability, but also that it can have aggravating aspects if it undermines capacity for rehabilitation and enhances future dangerousness,” and referenced *Atkins*, along with *Cullen v. Pinholster* (2011) and *Wong v. Belmontes* (2009).

In 2002, Michael DeWayne Smith of Oklahoma City, Oklahoma, was arrested and charged with two counts of murder and eventually sentenced to death. He later claimed that trial counsel had erred because they did not present evidence that he was intellectually disabled due to prior drug abuse and organic brain damage. (The initial trial judge concluded that *Atkins* did not apply since two of the three IQ tests the defendant took resulted in a score above that defined to be intellectually disabled. According to Oklahoma statute, any IQ test score 76 or above disqualifies one from being classified as intellectually disabled, and Smith had scores of 71, 76, and 79). The Oklahoma Court of Criminal Appeals rejected his claim in part because the court believed that such evidence had a “doubled-edged quality” that “might bolster a conclusion that the defendant represents a continuing threat to society.” (*Smith II* 2010).

And in 2014, Brandon Daniel was convicted of capital murder for killing a police officer and sentenced to death. The defense pointed out that Daniel had no previous history of violence, and several psychiatrists argued that Daniel committed his crime without forethought or deliberateness and that his actions were the result of severe depression, “a combination of the Xanax, his upbringing, the alcoholism and drug addiction of his parents, his traumatic brain injury to his frontal lobe, and his long history of impulse control” (*Daniel v. Texas* 2016). In particular, without drugs he would be unlikely to do anything violent. One psychologist noted that Daniel should respond well to a structured environment, that “many of the things that influenced him and

pushed him into a negative state during his childhood, youth, [and] early adulthood will be absent from him [in prison], particularly the substance abuse problems.” Nonetheless, the appellate court concluded that, because Daniel might be able to illegally access drugs while in jail, the evidence supported the jury’s ruling that he posed a future danger. That is, the court appears to have concluded that, in contrast to the testimony of the experts, Daniel’s addiction enhanced his potential for future dangerousness, even while he was in prison.

Interpreting Neuroscience Data

I recognize that these few G2i cases cannot support a definitive conclusion that neuroscience data are interpreted in light of the prejudices of the deciders of fact. Nonetheless, they do suggest that neuroscience data are largely being used, at least in our database, not to mitigate but to justify what juries and judges already have decided is the truth. It is noteworthy that for 5 out of the 34 cases in which the defendant was sentenced to less than LWOP (or a numerical number of years in prison that would effectively guarantee LWOP) were decided in favor of the prisoner (a 14.7% success rate), but only 2 out of the 21 death or LWOP cases were (a 9.5% success rate). From a purely clinical perspective, one would expect the success rates to be the same across sentencing types, for brain damage resulting in intellectual disability should be mitigating, per *Atkins*. However, this is not what we find. Instead, the longer the sentence, the less likely it is that evidence of brain damage will change the outcome.

Moreover, state laws currently determine whether someone is considered intellectually disabled such that the person cannot be executed. As noted in *ex parte Cathey* (2014): “Although psychology and psychologists inform the factual decision, they do not determine whether an inmate is exempt from execution under *Atkins*. We must apply our own judgment on the ‘appropriate ways’ to enforce the ultimately legal prohibition on executing mentally retarded offenders. *Atkins* did not conclude that there was a national consensus concerning the definition of mental retardation; rather, the Supreme Court concluded that there was a national consensus against execution of those offenders who fit within a given state’s definition of mental retardation, while permitting the states to continue to refine the contours of that definition in their own ways.” (See also *U.S. v. Wilson* 2013). And state laws vary, with several (e.g., Kentucky, Maryland, New Mexico, Ne-

braska, North Carolina, South Dakota, Tennessee, and Washington) holding that an IQ score of 70 is the maximum score one can receive to be declared intellectually disabled *in addition to* having problems with adaptive functioning and its appearance before the age of majority. That is, these states have made the bar for being declared intellectually disabled stricter since *Atkins*, instead of recognizing the inherent fuzziness of group-to-individual considerations.

I submit that the biggest challenge with the legal system is being able to utilize neuroscience data appropriately, when it is relying on experts to articulate test results but expecting judges and juries to interpret what these results mean. Because lay people do not have the expertise to do this, they fall back on what they already hold to be true about the defendant and then misinterpret the data to fit their beliefs. This can lead to some odd and unfortunate results.

Ex parte Cathey (2014) presents a good case in point. After being sentenced to death for capital murder, Eric Dewayne Cathey filed a writ of habeas corpus, claiming that he was ineligible for the death penalty because he was intellectually disabled. The Texas Court of Criminal Appeals denied the writ, noting that his score on the WAIS-R was only “borderline intellectual functioning.” The judge remarked that “courts should not become so entangled with the opinions of psychiatric experts as to lose sight of the basic factual nature of the *Atkins* inquiry: Is this person capable of functioning adequately in his everyday world with intellectual understanding and moral appreciation of his behavior wherever he is? Or is he so intellectually disabled that he falls within that class of mentally retarded inmates who are exempt from the death penalty?” However, the same judge also used irrelevant data to argue that Cathey was not intellectually disabled. He pointed out that on a word association test, Cathey scored in the “high average range” of 81%. And on the Trails B test, he scored in the 75th percentile. While it might seem reasonable to conclude that tests of word association and attentional capacity (the Trails B test) correlate with intellectual disability, such is not the case. One can have a low IQ, be impaired in everyday functioning, and still be verbally fluent or able to pay attention to multiple stimuli. If an expert were allowed to testify regarding how to interpret the data, such mistakes could perhaps be avoided. Indeed, it might be better to be “entangled with the opinions of psychiatric experts.” (I should point out

that, while concurring with the decision, Judge Price wrote that he disagreed “with the Court’s decidedly non-diagnostic approach to evaluating the adaptive-deficits prong of the standard for determining intellectual disability *vel non*.”)

And these mistakes can run both ways. In 1997, Jerry Ray Davidson of Dickson County, Tennessee, a previously convicted sex offender, was tried for first-degree premeditated murder, including mutilation of the body after death and aggravated kidnapping, for which he received the death penalty plus 20 years in prison. In contradistinction to other cases that warned of the double-edged dangers of using brain abnormalities, however, the Supreme Court of Tennessee found in favor of Davidson, concluding that his counsel was ineffective because they “possessed evidence that their client suffered from severe lifelong cognitive impairments and personality disorders and that he was predisposed to sexual violence. However, during the defendant’s trial, defense counsel presented the jury with no psychological mitigation evidence and only cursory social history mitigation evidence” (*Davidson v. State* 2014). The Tennessee court seized on brain data which showed that Davidson had some “mild atrophy” in his brain and a “slight excess of asynchronous slowing in all quadrants” of his EEG as information that defense counsel did not present, as well as prison medical records that “reveal Mr. Davidson to be a very mentally disturbed individual.” Instead of using neuroscience data and psychiatric histories, defense counsel had opted to rely solely on personal testimony from Davidson’s friends, family, and co-workers. The court remanded the case for a new sentencing hearing; and ultimately Davidson’s sentence was commuted to LWOP plus 20 years.

However, defense attorneys had procured his previous medical records from the Tennessee prison system when Davidson was previously incarcerated, which showed that he “had been evaluated and unsuccessfully treated during several previous incarcerations for sex crimes and that he had fantasies of raping and hurting women”. These records also contained “warnings from those who had tried and failed to rehabilitate Mr. Davidson that he would be a serious danger to the public following his release from prison.” In addition, an early examination by a forensic psychiatrist indicated that Davidson is not “fully aware of the social consequences of all his behavior,” and “when asked why he had been admitted to a mental hospital as a teenager, Mr. Davidson said it was because he ‘liked to grab at girls’ puss-

ies’.” Davidson recounted that “his favorite activities as a child were ‘raping girls, movies, fishing, and skating.’” A later report also noted that he “has demonstrated a remarkable lack of insight into his behavior, expresses no feelings of guilt about the incidents, and expects to behave similarly when he is released.” The Tennessee Court also recounts the “alarming content” of Davidson freely admitting to over 100 rapes (though other psychiatrists claimed he exaggerated the number). The prison psychologist concluded that, “Mr. Davidson does not appear to have the psychological resources to effectively deal with an unsupervised placement in the free world community.”

Despite all these warnings, Davidson had to be released from his previous prison terms once his sentence was served. In response, a clinical psychologist wrote to the Tennessee Board of Pardons and Paroles, worried that Davidson “may be inclined to commit rape/murder.” Moreover, staff members in the prison openly expressed concerns that “he would likely be a very real threat to others when released.” A final psychological evaluation concluded that he “may have difficulties controlling his impulses and may act out without thought of his consequences to his behavior,” and that he “seems to accept violence as a way of maintaining control over women.” Appeals counsel engaged additional experts, who concurred with previous evaluations and concluded that Davidson’s reasoning skills were at the level of a nine or ten year old, that he has “deficits in discriminating between actual information and distorted approximations,” and that he was “unable to accurately interpret and respond to social cues.”

Shockingly, the court was “not persuaded by the State’s argument that opening the door to Mr. Davidson’s mental health records would be a two-edged sword that hurt Mr. Davidson at sentencing more than it helped” because the jury already knew about his extensive history of sexual violence against women. It argued that instead of foregoing presenting any of Davidson’s medical and mental health history, which would keep all the previous information away from the jury, “Mr. Davidson’s counsel could have used their client’s threatening statements about raping women to illustrate his cognitive deficiencies,” and that the post-conviction experts could explain “how Mr. Davidson’s violent history and disturbing fantasies revealed the woundedness of his ramshackle mind.”

Again (and leaving aside the question of

whether death should ever be considered a just response by contemporary society), the court’s suggestions reveal a misunderstanding of the data. While it is true that Davidson very likely has “cognitive deficiencies” and a wounded and “ramshackle mind,” it is also the case that all of these evaluations point directly to the inevitability of future violence. These are, in fact, all aggravating factors, no matter how the court tries to spin them. I agree that they might “very well have helped the jury understand how Mr. Davidson could have committed “such horrendous acts’;” however, sentencing decisions do not turn on whether a jury understands the psychology of the defendants. More importantly, as Scalia accurately noted in his *Atkins* dissent, they are not solely about degree of culpability. Potential for rehabilitation and determination of future dangerousness are both important considerations. Defense counsel was right in its decision to try to humanize Davidson by having those who love him testify about what is good about him instead of having experts explain that he has a compulsion to hurt women that he likes and that no one can fix.

Again, we see a court interpreting data to fit its preconceived notions of what should happen, instead of what experts believe that data actually say. As I hope you can see through the cases I have briefly discussed, assuming that the meaning of neuroscience data should be left up to lay people — whether state legislators, judges, or jurists — is a disastrous way to manage expert G2i testimony in criminal cases. It can penalize some wrongly and can belie punishments for others. We need a better way.

References

- Alicke M D (2000) Culpable control and the psychology of blame. *Psychological Bulletin*, 126: 556–574.
- Alicke M D (2008) Blaming badly. *Journal of Cognition and Culture*, 8: 179–186.
- Alicke, M D, Buckingham J, Zell E, Davis T (2008) Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin*, 34: 1371–1381.
- Alicke M D, Rose D, Bloom D (2008). Causation, norm violation, and culpable control. *Journal of Philosophy*, 108: 670–696.

- Anwar S, Bayer P, Hjalmarsson R (2012) The impact of jury race in criminal trials. *The Quarterly Journal of Economics*, 127: 1017–1055.
- Atkins v. Virginia*, 536 U.S. 304 (2002).
- Barnett B (2006) *Medea* in the media: Narrative and myth in newspaper coverage of women who kill their children. *Journalism*, 7: 411–432.
- Benforado A (2010a) Frames of injustice: The bias we overlook. *Indiana Law Journal*, 85: 1332–1378.
- Benforado A (2010b) The body of the mind: Embodied cognition, law, and justice. *St. Louis University Law Journal*, 54: 1185–1216.
- Benforado A, Hanson J (2005) The costs of dispositionism: The premature demise of situationist law and economics. *Maryland Law Review*, 64: 24–84.
- Benforado A, Hanson J (2008a). The great attributional divide: How divergent views of human behavior are shaping legal policy. *Emory Law Journal*, 57: 311–338.
- Benforado A, Hanson J (2008b). Naïve cynicism: Maintaining false perceptions in policy debates. *Emory Law Journal*, 57: 499–574.
- Benforado A, Hanson J (2008c). Legal academic backlash: The response of legal theorists to situationist insights. *Emory Law Journal*, 57: 1087–1146.
- Benforado A, Hanson J, Yosifon D (2004) Broken scales: Obesity and justice in America. *Emory Law Journal*, 53: 1645–1806.
- Berman D H, Hafner C D (1989) The potential for artificial intelligence to help solve the crisis in our legal system. *Communications of the ACM*, 32: 928–938.
- Brown v. State*, W L 5780718 (Tenn. Appl., 2015).
- Bruner J S (2003) *Making Stories, Law, Literature, and Life*. Boston, MA: Harvard University Press.
- Buckholtz J W, Faigman J W (2015) Promises, promises for neuroscience and law. *Current Biology*, 24: R861–R867.
- Button K S, Ioannidis J P A, Mikrysz C, Nosek B A, Flint J, Robinson E S J, Munafò M R (2013) Power failure: Why small sample size undermines the reliability of neuroscience. *Nature: Neuroscience*, doi:10.1038/nrn3475.
- Carter v. Bradshaw*, WL 1394059 (N.D. Ohio, E. Div., 2016).
- Catley P, d Claydon L (2015) The use of neuroscientific evidence in the courtroom by those accused of criminal offenses in England and Wales. *Journal of Law and the Biosciences*, 2: 510–549.
- Chandler J A (2015) The use of neuroscientific evidence in Canadian criminal proceedings. *Journal of Law and the Biosciences*, 2: 550–579.
- Chen R, Hanson J (2004a) Categorically biased: The influence of knowledge structures on law and legal theory. *Southern California Law Review*, 77: 1103–1254.
- Chen R, Hanson J (2004b) The illusion of law: The legitimating schemas of modern policy and corporate law. *Michigan Law Review*, 103: 1–149.
- Chesney-Lind M (1999) Media misogyny: Demonizing “violent” girls and women. In *Making Trouble: Cultural Constructions of Crime, Deviance and Control*, eds J Ferrell, N Websdale, 114–140. New York: Walter de Gruyter, Inc.
- Cullen v. Pinholster*, 131 S.Ct. 1388 (2011).
- Daniel v. Texas*, WL 519836 (Tex. Crim. App. 2016).
- Davidson v. State*, 453 S.W.3d 386 (S.C. Tenn. 2014).
- Davis P C, Wippold F J II, Cornelius R S, Aiken A H, Angtuaco E J, Berger K L, Broderick D F, Brown D C, Douglas A C, McConnell C T Jr, Mechtler L L, Prall J A, Raksin P B, Roth C J, Seidenwurm D J, Smirniotopoulos J G, Waxman A D, Coley B D (2012) Expert panel on neurologic imaging. ACR Appropriateness Criteria® head trauma. Reston, VA: American College of Radiology (ACR). <http://www.guideline.gov/content.aspx?id=37919>.
- de Kogel C H, Westgeest E J M C (2015) Neuroscientific and behavioral genetic information in criminal cases in the Netherlands. *Journal of Law and the Biosciences*, 2: 580–605.
- Denno D W (2015) The myth of the double-edged sword: An empirical study of neuroscience evidence in criminal cases. *Boston College Law Review*, 56: 493–551.
- Denno D W (2017). How prosecutors and defense attorneys differ in their use of neuroscience evidence. *Fordham Law Review*, 85: 453–479.
- Devoe v. Davis*, WL 5408169 (W.D. Texas 2016).
- Devoe v. State*, 354 S.W.3d 457, 462–66 (Tex. Crim. App. 2011).
- Eberhardt J L, Davies P G, Purdie-Vaughns V J, Johnson S L (2006) Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science*, 17: 383–386.
- Ex parte Cathey*, 451 S.W.3d 1 (Tex. Crim. App. 2014).

- Faigman D L, Monahan J, Slobogin C (2014) Group to individual (G2i) inference in scientific expert testimony. *University of Chicago Law Review*, 81: 417–480.
- Farr K A (1997) Aggravating and differentiating factors in the cases of white and minority women on death row. *Crime and Delinquency*, 43: 260–278.
- Gaudet L M, Marchant G E (2016) Under the radar: Neuroimaging evidence in the criminal courtroom. *Drake Law Review*, 64: 577–661.
- Goff P A, Eberhardt J L, Williams M J, Jackson M C (2008) Not yet human: Implicit knowledge, historical dehumanization, and contemporary consequences. *Journal of Personality and Social Psychology*, 94: 292–306.
- Hall v. Florida*, 572 U.S. (2014).
- Hanson J (ed.) (2012). *Ideology, Psychology, and Law*. New York, NY: Oxford University Press.
- Hanson J, McCann M (2008) Situationist torts. *Loyola Law Review*, 41: 1345–1454.
- Hanson J, Yosifon D (2003) The situation: An introduction to the situational character, critical realism, power economics, and deep capture. *University of Pennsylvania Law Review*, 152: 129–346.
- Hanson J, Yosifon D (2004) The situational character: A critical realist perspective on the human animal. *Georgetown Law*, 93: 1–179.
- Hardcastle V G (2017) Brain images in the courtroom: An analysis of recent appellate decisions in criminal cases. In *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*, eds G Caruso, O Flanagan, 311–332. New York: Oxford University Press.
- Hardcastle V G, Kitzmiller M K, Lahey S (2018) The impact of neuroscience data in criminal cases: Female defendants and the double-edged sword. *New Criminal Law Review*, 21: 291–315.
- Hardcastle V G (forthcoming 2018) Diversion courts, traumatic brain injury, and American vets. In *Neuro-Interventions and the Law: Regulating Human Mental Capacity*, eds N Vincent, T Nadelhoffer, A McCay. New York: Oxford University Press.
- Hergovich A, Schott R, Burger C (2010) Biased evaluation of abstracts depending on topic and conclusion: Further evidence of a confirmation bias within scientific psychology. *Current Psychology*, 29: 188–209.
- Huckerby J (2003) Women who kill their children: Case study and conclusions concerning the differences in the fall from maternal grace by Khoua and Andrea Yates. *Duke Journal of Gender, Law, and Policy*, 10: 149–172.
- Kahneman D (2011) *Thinking Fast and Slow*. New York: Farrar, Strauss, and Giroux.
- Kang J, Bennett M, Carbado D, Casey P, Dasgupta N, Faigman D, et al. (2012) Implicit bias in the courtroom. *UCLA Law Review*, 59: 1124–1186.
- Kunda, Z. (1999). *Social Cognition: Making Sense of People*. Cambridge, MA: The MIT Press.
- Levinson, J. D. (2007). Forgotten racial equality: Implicit bias, decision-making, and misremembering. *Duke Law Journal*, 57, 345–424.
- Levinson, J. D., and Young, D. (2010). Different shades of bias: Skin tone, implicit racial bias, and judgments of ambiguous evidence. *West Virginia Law Review*, 112, 307–350.
- Levinson, J. D., Cai, H., and Young, D. (2010). Guilty by implicit racial bias: The guilty/ not guilty implicit association test. *Ohio State Journal of Criminal Law*, 8, 187–208.
- Levinson, J. D., Smith, R. J., and Young, D. M. (2014). Devaluing death: An empirical study of implicit racial bias on jury-eligible citizens in six death penalty states *New York University Law Review*, 89, 513–581.
- Levinson, J.D. (2007). Forgotten racial equality: Implicit bias, decision making, misremembering. *Duke Law Journal*, 57, 345–424.
- Lewin, J.L. (1998). The genesis and evolution of legal uncertainty about “reasonable medical certainty.” *Maryland Law Review*, 57, 397–406.
- Lucas v. State*, 274 Ga. 640, 555 S.E.2nd 440 (S.C. Georgia, 2001).
- Lucas v. Warden*. 771 F.3d 785 (U.S. Ct. of App., 11th Cir 2014).
- Maccoun, R.J. (1998). Bias in the interpretation and use of research results. *Annual Review of Psychology*, 49, 259–287.
- Matlin, M.W. (2004). Pollyanna principle. In R.F. Pohl (Ed.), *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgment and Memory* (pp. 255–273). New York: Psychology Press.
- McCarty, L.T. (1977). Reflections on “Taxman”: An experiment in artificial intelligence and legal reasoning. *Harvard Law Review*, 90, 837–893.
- Moore v. Texas*, 581 U.S. ____ (2017).
- Moriarty, J. C. 2008. Flickering neuroadmissibility: Neuroimaging evidence in U.S. Courts. *Behavioral Sciences and the Law*, 26, 29–49.
- Morse, S. J. (2012). Neuroimaging evidence in law: a plea for modesty and relevance. In: J. R. Simpson (Ed.), *Neuroimaging in forensic psychia-*

- try: *From the clinic to the courtroom* (pp. 341–358). Chichester: Wiley-Blackwell.
- Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2, 175–200.
- Oregon v. Agee*, CC 09C41224; SC S059530 (Sup. Ct. Ore, 2015).
- Oregon v. Davis*, Case No. 020633788 (Cir. Ct. Ore., 2016).
- Oswald, M.E. and Grosjean, S. (2004). Confirmation bias. In R.F. Pohl (Ed.), *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgment and Memory* (pp. 79–96). New York: Psychology Press.
- Payne, B. K. (2006). Weapon bias: Split-second decisions and unintended stereotyping. *Current Directions in Psychological Science*, 15, 287–291.
- Penry v. Lynaugh*, 492 U.S. 302 (1989).
- People v. Nash*, WL 4041718 (Cal. Appl. 5th, 2015).
- Reynolds, C. (2013). Implicit bias and the problem of certainty in the criminal standard of proof. *Law and Psychology Review*, 37, 229–248.
- Roberson v. Stephens*, WL 4720270 (U.S. Ct. of App., 5th Cr., 2015).
- Salisbury, E.J., and Van Voorhis, P. (2009). Gendered pathways: A quantitative investigation of women probationers' path to incarceration. *Criminal Justice and Behavior*, 36, 541–566.
- Sarat, A. (1993). Research on the death penalty: Speaking of death: Narratives of violence in capital trials. *Law and Society Review*, 27, 19–58.
- Shen, F.X. (2017). Neuroscientific evidence as instant replay. *Journal of Law and the Biosciences*, 2, 343–349.
- Smith II*, 245 P.3 1233 (Okla. Crim. Appl. 2010).
- Smith v. Duckworth*, 824 F.3d 1233 (U.S. Ct. of Appl. 10th Cir, 2016).
- State v. Norzagaray*, WL 1903591 (App. Ct. Ariz. Div 2, 2015).
- Susskind, R. (1986). Expert systems in law: A jurisprudential approach to artificial intelligence and legal reasoning. *Modern Law Review*, 49, 168–194.
- Tenorio, E.H. (2011). Critical Discourse Analysis, an overview. *Nordic Journal of English Studies*, 10, 183–210.
- Treadway, M.T., and Buckholtz, J.W. (2011). On the use and misuse of genomic and neuroimaging science in forensic psychiatry: Current roles and future directions. *Child and Adolescent Psychiatry Clinics of North America*, 20, 533–546.
- Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Uhlmann, E. L., and Cohen, G. L. (2007). “I think it, therefore it’s true”: Effects of self perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, 104, 207–223.
- United States v. Wilson*, 922 F.Supp.2d 334 (E.D.N.Y 2013).
- Van Dijk, T.A. (2008). *Discourse and Context: A Sociological Approach*. New York, N.Y.: Cambridge University Press.
- Watson, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 129–140.
- Weiss, G., and Wodak, R. (Eds.) (2003). *Critical Discourse Analysis: Theory and Interdisciplinarity*. New York: Palgrave MacMillan.
- Wilczynski, A. (1991). Images of women who kill their infants: The mad and the bad. *Women and Criminal Justice*, 2, 71–88.
- Williams v. Kelley*, 581 U.S. 16–8923 16A1045 (2017).
- Williams v. State*, 51 S.W.3d 290 (Ark. 2007).
- Wodak, R. and Meyer, M. (2001). Critical Discourse Analysis: History, agenda, theory, and methodology. In R. Wodak and M. Meyer (Eds.), *Methods of Critical Discourse Analysis* (pp. 1–33). London: Sage.
- Wodak, R., (2008). Introduction: Discourse studies – important concepts and terms. In R. Wodak and M. Krzyzanowski (Eds.), *Qualitative Discourse Analysis in the Social Sciences* (pp. 1–29). New York: Palgrave MacMillan.
- Wong v. Belmontes*, 130 S. Ct. 383 (2009).

Invited paper

LOOPING EFFECTS OF NEUROLAW, AND THE PRECARIOUS MARRIAGE BETWEEN NEUROSCIENCE AND THE LAW

Toma Strle, Olga Markič*

University of Ljubljana, Faculty of Education, Kardeljeva ploščad 16, 1000 Ljubljana,

e-mail: toma.strle@pef.uni-lj.si

*University of Ljubljana, Faculty of Arts, Aškerčeva 2, 1000 Ljubljana

e-mail: olga.markic@ff.uni-lj.si

Abstract:

In the following article we first present the growing trend of incorporating neuroscience into the law, and the growing acceptance of and trust in neuroscience's mechanistic and reductionistic explanations of the human mind. We then present and discuss some studies that show how nudging peoples' beliefs about matters related to human agency (such as free will, decision-making, or self-control) towards a more deterministic, mechanistic and/or reductionistic conception, exerts an influence on their very actions, mentality, and brain processes. We suggest that the neuroscientific view of the human mind exerts an influence on the very cognitive phenomena neuroscience falsely believes to be studying objectively. This holds especially when we consider the systematic integration of neuroscience into the public domain, such as the law. For, such an integration acts as a reinforcement of the public's and legal decision-makers' endorsement of and trust in neuroscience's view of human nature that further changes how people think and act. Such looping effects of neurolaw are probably inevitable. Accordingly, we should be aware of the scope of neuroscientific explanations and be careful not to overstate neuroscientific evidence and findings in legal contexts.

Key words: agency, belief, looping effects, natural and human kinds, moral responsibility, neurolaw, self-control.

1. The growing trend of marrying neuroscience¹ and the law

For several decades, the introduction and admittance of neuroscientific evidence in courts of the law have been on the rise. One of the first cases is John Hinckley's trial for the attempted assassination of Ronald Reagan (in 1981) where CT scans were presented with the intention of demonstrating John Hinckley's schizophrenia (Church, 2012). In *People v. Weinstein* (1992), Herbert Weinstein's defence introduced PET scans during his murder trial to argue that a brain cyst had been responsible for him murdering his wife (Jones, Marois et al., 2013). More recently, 2005 saw the defence lawyer of Grady Nelson — who had brutally murdered his wife — argue for mitigated punishment on the grounds that QEEG evidence of relevant brain abnormalities allegedly led to his criminal actions (Jones, Wagner et al., 2013). Further on in 2007, the defence of Peter Braunstein — who had been sentenced to eighteen years to life for kidnapping, sexual abuse, and robbery — presented MRI evidence to argue that schizophrenia had prevented him from controlling his violent impulses (Church, 2012).

¹ In the article, we refer to the part of neuroscience that deals with mental phenomena or provides explanations of them — be it on the level of molecules, neurons or networks — and is often dubbed cognitive neuroscience. But it has to be pointed out that if neuroscience were solely about neurobiology, it would not be of much interest to sciences of the mind or the law.

Moreover, Farahany's analysis of the use of neurobiological data as evidence in the U.S. courts clearly reveals the growing trend of incorporating neuroscience into the law:

Over 1585 judicial opinions issued between 2005 and 2012 discuss the use of neurobiological evidence by criminal defendants to bolster their criminal defence. In 2012 alone, over 250 judicial opinions—more than double the number in 2007—cited defendants arguing in some form or another that their 'brains made them do it'. (2016, p. 486)²

What is striking about Farahany's analysis is that around 60% of the analysed cases were non-capital offences such as robbery, fraud, and drug trafficking.

Furthermore, the use of neuroscientific evidence in courts is not only on the rise in the U.S. and Common Law countries; in fact, it has been utilised by defendants in other (and non-Common Law) countries: The Netherlands, Singapore, Canada, England, Wales, India, Italy, etc. (Catley, 2016; Church, 2012). Italy's court of appeals, for example, is cited by Church (2012) as the first European court to use brain imaging scans (besides genetic data) to mitigate the defendant's murder sentence, "finding that the evidence indicated an unavoidable propensity toward violence" (Church, 2012, p. 1828). Moreover, Catley and Claydon (2015) identified 204 cases in England and Wales between 2005 and 2012 in which neuroscientific evidence was used. Their analysis shows a notable increase in the use of neuroscientific evidence by those accused of criminal offences: in the period of 2005-2008 the annual average was just under 17 cases; in the period of 2009-2012 this figure doubled (just over 34 cases).

All in all, the rapidly growing understanding of the human brain, advances in non-invasive functional imaging techniques, and the growing popularity of neuroscience in and out of courts show that neuroscience is on its way to become an integral part of the law. In addition, the growing trend of merging neuroscience and the law points to the conclusion that neuroscientific explanations of the human mind and behaviour are becoming more accepted and trusted by legal decision-makers, the public, and the (neuro)scientific community.

Aside from the trend itself, the trust in neuroscientific explanations and descriptions of (criminal)

behaviour and mentality is revealed by several studies and trials. These studies show, for instance, how spelling something out in the language of physiology (e.g., neuroscience), or presenting physiological rather than psychological data, alters people's judgments about culpability, moral responsibility, free will, etc. Consider these illustrations: research by Gurley and Marcus (2008) demonstrates that judges and juries are more inclined to accept the insanity defence when neuroscientific evidence is presented; research by Monterosso, Royzman, and Schwartz (2005) shows that people judge behaviour of others less harshly when explained in light of physiology rather than psychology; and, research by Nahmias, Coates, and Kvaran (2007) reveals that the way a deterministic scenario (the world and actions of hypothetical agents) is described has a significant impact on people's judgments about free will and moral responsibility — the study shows how neuroscientific descriptions prompt people to ascribe less free will and moral responsibility to agents than do psychological descriptions of the same deterministic scenarios. Note that one of the points of Nahmias et al. (2007) is that, even though the scenarios are described in a deterministic manner, it is neuroscience's *reductionistic* language which induces people to ascribe less free will and moral responsibility to agents, not the determinism itself.

Moreover, the trust in neuroscientific evidence is also clearly visible in, for instance, the trial of Grady Nelson, mentioned earlier. In the trial, two out of six jurors (from twelve altogether), who voted against the death penalty, reported that QEEG evidence overturned their judgment (the two jurors had supported the death penalty before the evidence was presented). Miller (2010) reports:

But two of those who voted against it told *The Miami Herald* that the QEEG evidence was influential: *Delores Cannon, a hospital secretary, said she leaned toward death until the technology was presented. "But then when it came in, the facts about the QEEG, some of us changed our mind," she said. John Howard, an airport fleet services worker, said he, too, was ready to recommend death. "It turned my decision all the way around," Howard said the QEEG [sic]. "The technology really swayed me . . . After seeing [sic] the brain scans, I was convinced this guy had some sort of brain problem."* (para. 6)

² Her analysis included cases where neuroscientific and/or behavioural genetics evidence were discussed.

What is more, neuroscientific evidence is being admitted in courts beyond what the neuroscientific community would — *at the moment* — deem appropriate. Church (2012) reports how, in 2008, India became the first country to convict a criminal (named Sharma) for murder on the basis of brain fingerprinting technology (with two more offenders in six months).³ The response of the majority of commentators was, however, highly sceptical, “believing that without more proof of the technology’s reliability, the court’s willingness to rely on it was unsoundly premature” (Church, 2012, p. 1827; footnotes omitted by authors). Nonetheless, the courts did avail themselves of the technology, and it is hard to see how the trend will not continue. And, even though lie-detection technology has not been to our knowledge admitted as evidence in the U.S. or Europe, the defence tried to introduce it in *United States v. Semrau* in 2010 (Church, 2012). And yet, as Church points out, the judge, excluding the evidence, noted that

he could foresee a time in the future when this evidence could, in fact, be admissible ... The judge based this opinion on an assumption that “fMRI-based lie detection [will eventually] undergo further testing, development, and peer review, improve upon standards controlling the technique’s operation, and gain acceptance by the scientific community for use in the real world.” (pp. 1837–1838; footnotes omitted by authors)

In fact, neuroscientific evidence of and findings about brain damage and functional abnormalities — allegedly illuminating mental states of offenders and revealing the causes of defendants’ actions — have not only been used in particular cases for mitigating punishment, challenging defendants’ competency and the effectiveness of defence counsel, as Farahany’s (2016) analysis reveals, but have already been used as a basis for decisions pertaining to more general questions about (criminal) responsibility. In *Roper v. Simmons* (2005), for instance, the advocates of the abolition of death penalty for adoles-

cents argued that neuroscientific findings demonstrate that brains of adolescents are not sufficiently developed and adolescents thus insufficiently rational and not as responsible for their actions as adults (as cited in Morse, 2006; see pages 406 – 410 for an illuminating analysis of the case). In its decision not to execute Simmons (who had been seventeen years old when he committed the murder), the U.S. Supreme Court did not state neuroscientific findings as the reason for the abolition of death penalty for adolescents (Morse, 2006). However, five years later the U.S. Supreme Court did cite neuroscientific findings in *Graham v. Florida* to categorically exclude juveniles from life without the possibility of parole in non-homicide cases (Morse, 2015; see also *Miller v. Alabama*, 2012, also in Morse, 2015).

Last but not least, neuroscience has also been discussed in the context of understanding the neural basis of judicial decision-making and biases that can accompany it (e.g., among eyewitnesses, juries, judges, and parole officers), or in regard to creating direct brain intervention therapies for criminals (Jones, Marois et al., 2013). At the extreme, neuro-based prediction is being discussed as a means for predicting the future behaviour of offenders — also in regard to decisions regarding parole, sentencing, bail, and death penalty (Jones, Marois et al., 2013), or future re-arrest (e.g., Aharoni et al., 2013). One could even speculate that neuroscience could, in a foreseeable future, be used for predicting future criminal actions and thus open up the possibility of punishing the “crime of thought”.

The examples, studies, and evidence presented so far indicate that neuroscience is not only on the way to becoming an integral part of the law and that its explanations of the human mind and behaviour are becoming more accepted, but that neuroscience has most probably changed — and may continue to do so — our views and intuitions of other’s and our own agency, personhood, decision-making, and responsibility — concepts that are at the core of the law. As Farah (2012) succinctly puts it:

Neuroscience does not merely give us new tools to be used to the benefit or detriment of humanity; it gives us a new way of thinking about humanity. ... Neuroscience provides an alternative perspective, from which human behavior can also be understood as the result of physical causes. ... [T]he applications of neuroscience ... will provide many reminders that our minds are, at root, physical mechanisms. By making people part of the clock-

³ Brain fingerprinting is an EEG technique that supposedly detects concealed information in the brain by detecting specific ERPs; thus, it potentially enables recognition of concealed memories. It is sometimes stated to be a lie-detection technology, although it measures cognitive information processing (Farwell, 2012). FMRI lie-detection, on the other hand, is being developed to determine brain activity when a person is lying (see Farah et al., 2014).

work universe, neuroscience challenges many assumptions about morality and personhood. ... We do not say, “But he had no choice—laws of physics made him do it!” However, as the neuroscience of personality, decision making, and impulse control begins to offer a more detailed and specific account of the physical processes leading to irresponsible or criminal behavior, the deterministic viewpoint will probably gain a stronger hold on our intuitions. (pp. 585–586)

2. Looping effects of neurolaw and objectivistic neuroscience

Admittedly, the (neuro)scientific community is relatively sceptical about the use of neuroscientific evidence in courts *at the moment*, as well as being critical of the *current* reach of neuroscience (e.g., Jones, Marois et al., 2013; Jones, Wagner et al., 2013).⁴ However, the neuroscientific community does generally accept the supposition that neuroscience will, in the future, be able to elucidate the true workings of the human mind in a more-or-less mechanistic and reductionistic manner,⁵ thus providing the best, if not a complete, explanation of human behaviour. What is more, one cannot help but notice the seduction (as is nicely expressed by Farah’s quote above) of the common belief that the true (objective) science (i.e., neuroscience) of the human mind has finally been found (for such implicit and/or explicit view of neuroscience, see: Gazzaniga, 2008; Greene and Cohen, 2004; Jones, Marois et al., 2013; Jones, Wagner et al., 2013; Jones et al., 2014).⁶ Namely, a science that will purportedly allow us to *construct* a unified — mechanistic and reductionistic — model of human mentality and, at last, provide a solid explanation of human behaviour that is interesting to the law. However, we believe that such a view of human nature (and of human agency, decision-making, self-control, etc.), as endorsed by neu-

rosience, is quite incorrect (see Bennett and Hacker, 2003; Glannon, 2009; Kordeš, 2016; Markič, 2013; Morse, 2006; Morse, 2015; Strle, 2013; Strle, 2016; Uršič, Markič, and Ule, 2012; Varela, Thompson, and Rosch, 1991; Vörös and Markič, 2014).⁷

Since many have criticised neuroscience’s objectivistic and/or reductionistic view(s) of human mentality, we will point out another matter which is left out of most discussions about the relation of neuroscience and the law. We will argue (also on the grounds of empirical evidence) that the increasingly popular and accepted objectivistic, deterministic, and reductionistic view of human agency, decision-making, self-control, responsibility, etc. (i.e., phenomena important to the law), as generally endorsed by neuroscience, is changing how we think, act, and *brain* (pun intended). To put it differently, phenomena that neuroscience objectifies, and to which it applies reductionist explanations, change — at least to a certain degree — according to how they are described and conceptualised. As such, neuroscience’s view of the human mind exerts an influence on the very phenomena neuroscience claims to be studying objectively, and which the law uses for the purpose of determining culpability — a curious looping effect, indeed.⁸ This particularly holds when we consider the systematic integration of neuroscience into the public domain, such as the law (e.g., Gazzaniga, 2008; Greene and Cohen, 2004; Jones et al., 2014) — a trend which is also visible from the cases and literature presented in the first chapter of the article. For, such an integration acts as a reinforcement of the public’s and legal decision-makers’ endorsement of and trust in neuroscience’s view of human nature, thus further changing how people think and act. Consequently, this alters neuroscience’s very (future) findings.

But, if such looping effects do exist, this means that neuroscience cannot be taken at face value as *the true objective* science of the human mind and behaviour in legal contexts and elsewhere. At the end of the article, we will discuss in some more detail what this means or should mean for the precarious engagement of neuroscience and the law.

⁴ From the perspective of the reductionist and representationalist paradigm espoused by neuroscience, one should add.

⁵ In the article, we refer to methodological and epistemic reduction (especially the explanatory variant), but not ontological (for the distinction see Brigandt and Love, 2017; as well as Strle, 2013).

⁶ For instance, the recommendations submitted to the U.S. President’s Bioethics Commission by Jones et al. (2014) are — although critical and calling for further research — explicitly and implicitly quite optimistic about the potential reach of neuroscience in elucidating matters of mind and behaviour pertaining to the legal context.

⁷ By no means do we argue that conscious mental phenomena do not partly emerge from brain processes.

⁸ A similar idea, in the context of critical neuroscience, is expressed by Choudhury, Nagel, and Slaby (2009): “[R]epresentations of the objectified phenomenon influence the phenomenon itself and its subsequent study, and thus, the journey is a loop” (p. 65, text under figure 1).

3. Looping beliefs: neuroscience is not about natural but human kinds

Let us now present and discuss some studies that show how nudging peoples' beliefs and theories about matters related to human agency (such as free will, decision-making, or self-control) towards a more deterministic, mechanistic, and/or reductionistic conception exerts an influence on their very actions, mentality, and brain processes. Next, we will try to extrapolate these findings to the claim that the neuroscientific view of human mentality exerts an influence on the very phenomena neuroscience claims to be studying objectively (also through the growing acceptance of and trust in neuroscience's view of human nature, and its integration into the law). Lastly, we will present the distinction between natural and human kinds (Brinkmann, 2005; Brinkmann, 2014; Hacking, 1995), concluding that neuroscience is, in fact, researching human kinds, rather than the natural kinds it claims to study. These phenomena (i.e., human kinds) are intrinsically such that by changing our conception of them or our activities pertaining to them they themselves change. This is also supported by studies we will now present and discuss.

Vohs and Schooler's (2008) seminal study reveals that weakening peoples' belief in free will leads to changes in their moral actions (i.e., increased cheating). In the first experiment, participants first read one of the two passages from the book *The Astonishing Hypothesis* (Crick, 1994). In the anti-free will condition, participants read statements "claiming that rational, high-minded people—including, according to Crick, most scientists—now recognize that actual free will is an illusion, and also claiming that the idea of free will is a side effect of the architecture of the mind" (Vohs and Schooler, 2008, p. 50). In the control group, participants read a passage on consciousness that did not mention free will. Researchers then established that the anti-free will group reported weaker free-will beliefs (measured on a scale) than the control group. Then participants had to solve 20 arithmetic problems. They were told to press the space bar after the problem was presented on the computer screen to prevent the correct answer from appearing (the computer did, in fact, record space bar presses). The main result reveals that the anti-free will group cheated significantly more in comparison with the control group. But this experiment only revealed passive cheating.

In a second experiment, Vohs and Schooler tried to see whether their findings also hold for situa-

tions that enable active cheating. This experiment included five conditions. In all conditions, participants were given 15 reading-comprehension, mathematical, and logic and reasoning problems, being told they would earn \$1 for each correctly solved problem. In the first condition, participants did not read any statements before solving the problems, while participants in the second, before solving problems, read determinism-related statements (see below). In both, experimenters scored their answer sheets and paid them \$1 for each correctly solved problem, thus making cheating impossible in these two conditions. Conversely, cheating was made possible in the three additional conditions: before having to solve the aforementioned problems, participants either read free-will-related statements, determinism-related statements, or neutral statements.⁹ The first two were meant to manipulate participants' belief in free will; and, as hypothesised, participants in the free will condition reported stronger beliefs in free will than participants in the determinism or neutral conditions (again, measured on a scale). Participants then had to solve problems, being told they would get \$1 for each correctly solved problem. Contrary to the first two non-cheating conditions, the participants were not supervised by the experimenter: after some instruction, the experimenter looked at her mobile phone, said she had to attend a meeting, and instructed participants to carry on with the task. The participants themselves had to score the correctness of their solutions from answer sheets left in the room, pay themselves \$1 for each correct answer from an envelope nearby, and shred their answer sheets. These circumstances enabled participants to cheat since participants had to self-score and self-pay. After participants finished the experiment, the experimenters calculated the average payment per participant in each of the five conditions (they knew the initial amount of the money in the envelope for the three cheating conditions). To assess cheating, experimenters then compared payments between all conditions. The study revealed that participants in the self-scoring, self-payment determinism condition paid themselves far more

⁹ In the free will condition, they read statements such as "Avoiding temptation requires that I exert my free will" (Vohs and Schooler, 2008, p. 51). In the determinism condition, they read statements such as "A belief in free will contradicts the known fact that the universe is governed by lawful principles of science" (p. 51; note that the statement says *science*, not *nature*). In the neutral condition, they read statements unrelated to free will.

money (ca. \$11) than participants in other four conditions, which did not differ from each other (ca. \$7). Furthermore, the results revealed that the more participants agreed with statements of free will, the less money they paid themselves (on average).

All in all, Vohs and Schooler's findings "raise the genuine concern that wide-spread encouragement of a deterministic worldview may have the inadvertent consequence of encouraging cheating behavior" (p. 53). However, it is important not to over-interpret their findings: as they point out, the study measured only modest forms of immoral behaviour, so it is unknown whether the findings can be generalised to more significant actions; a deterministic viewpoint may also have favourable consequences, etc.¹⁰ On the other hand, their research does show that accepting a deterministic, mechanistic, and reductionistic view of human agency (Crick's book and the determinism supporting statements used in their research also endorse the latter two, at least indirectly) can undermine our sense of agency; thus, it exerts an influence on our (moral) actions.

Similar studies further demonstrate how the belief that we cannot exercise control over our behaviour and outer events leads to poorer performance (Bandura, 1989), that undermining belief in free will is linked to decreased willingness to help, as well as increased aggression (Baumeister, Masicampo, and DeWall, 2009), and induces impulsive and antisocial tendencies (Rigoni et al., 2012) which,¹¹ according to the authors, implies reduced self-control, etc. Rigoni et al. (2011) even showed that weakening people's belief in free will influences brain processes related to a basic motor level: the reduction of readiness potential, which is linked to voluntary motor preparation. As is succinctly put by Rigoni et al. (2011), such studies demonstrate the following:

[U]ndermining the idea that we are "the masters of our own houses" ("der Herr im eigenen Haus"; Freud, 1917, p. 11) presumably reduces the intentional effort we put into action. ... Putting less effort into an action might weaken our sense of agency for these actions and lead to a reduced feeling of responsibility. This reduced feeling of responsibility would very likely result in more careless and irresponsible behavior. ... From this perspective, we could hypothesize that dismissing the idea that people can control their own actions acts as a nonauthorship indicator, thereby decreasing people's sense of authorship. In sum, ... this suggests that abstract belief systems might have a much more fundamental effect than most people would expect. (p. 617)

Similar findings come from research on the ego-depletion effect, the claim that willpower or the capacity to exert self-control is limited in the way that exerting willpower on mental tasks leads to worse performance on subsequent mental tasks (Savani and Job, 2017). The research shows that the ego-depletion effect is quite dependent on peoples' implicit, explicit, and culturally accepted beliefs about the nature of self-control (e.g., Job, Dweck, and Walton, 2010; Savani and Job, 2017). For instance, Job et al. (2010) showed that lay theories about willpower moderate ego-depletion effects and affect self-regulation. Furthermore, Savani and Job (2017) demonstrated that the idea of ego-depletion is strongly dependent on people's cultural beliefs about the exertion of self-control. In their study, Indians, contrary to Americans, tended to believe that exerting willpower is energising and, consequently, exhibited the *reverse* ego-depletion effect: acts of self-control improved subsequent performance on mental tasks, contrary to the traditionally accepted belief that ego-depletion "is a pan-cultural phenomenon rooted in human biology" (Savani and Job, 2017, p. 14). The *reverse* ego-depletion effect was also exhibited by Americans who held similar beliefs about the exertion of self-control as Indians.

Savani and Job (2017) conclude their article with a statement that is quite in line with the idea of looping effects: "Peoples' cultural background and beliefs contribute considerably to what has been primarily viewed as a biological phenomenon" (p. 15). Namely, the phenomena under study (e.g., volition, agency, self-control, decision-making, etc., which are essential ingredients in modern law) are quite dependent on how they are conceptualised,

¹⁰ For instance, the study of Shariff et al. (2014) shows that reducing beliefs in free will reduces retribution since, according to the authors, believing in free will supports ascriptions of moral responsibility and thus retributive behaviour. While this could be seen as a positive effect from belief in a mechanistic and deterministic nature of human mind, it does not (for the purpose of the present article) matter whether the effects are positive or negative.

¹¹ In some of the trials presented at the beginning of the paper, the defence attorneys tried to argue for such antisocial tendencies solely on the grounds of neuroscientific evidence. Clearly, if beliefs play a major role in the co-determination of such tendencies (as this research shows), then referring solely to brain mechanisms and explanations in the attempt to demonstrate that defendants are not culpable for their actions is not adequate.

understood, and felt. But this means that our very understanding, beliefs, feelings, etc. (cultural or otherwise) about the phenomena exert an influence on the very phenomena that neuroscience claims (and/or hopes) to be researching objectively. That is, the phenomena under study are not in fact solid, unchanging objects that we just have to look at through the correct lens in order to understand and explain them properly; rather, they are phenomena that change according to our understanding and conceptualisation of them. As such, they are phenomena that can be influenced by the very conceptualisation a science applies to them — even more so if accepted by the public.¹² But, if we accept such looping effects to exist, then we must also accept the claim that those sciences which research conscious mentality (such as neuroscience) exert an influence on their very (future) findings.

This holds especially when considering the systematic integration of neuroscience into the public domain, such as the law. For, the growing trend of admitting neuroscientific evidence in courts and accepting the findings and theories of neuroscience in legal contexts acts as a reinforcement of the public's and legal decision-makers' endorsement of and trust in neuroscience's objectivistic and reductionistic view of human nature. (If nothing else, the law and its practices are to a certain degree representative of the public's beliefs and values.) And, the more the public endorses a reductionistic view of human nature, the more neuroscientific evidence, findings, and theories are taken into consideration in legal contexts when determining culpability, as well as being possible grounds for changing legal systems. This is another looping effect, one that not only reinforces the neuroscientific image of human nature, but one which also exerts further influence on peoples' actions, mentality, and brain processes.¹³ Admittedly, the latter claim holds only if the pre-

sented empirical findings can be extrapolated to the society at large, or at least to a part of it. This is indirectly supported by the growing trend of integrating neuroscience into the law, as presented in the first part of the article.

As a consequence of the described looping effects, neuroscience that deals with cognitive phenomena cannot be claimed to be an objective science of the mind. Indeed, cognitive phenomena are, *at their core*, social in nature: they are embedded in the social world, values, and practices (*as is neuroscience itself*), and thus are co-constituted by persons with bodies and values rather than merely by the brains of individuals.¹⁴

According to Hacking (1995),¹⁵ mental phenomena, and thus phenomena interesting to the law, are mostly not natural kinds (as neuroscience would like to have it). Instead, they are human kinds that, contrary to natural kinds, have the following characteristics, as is succinctly summarised by Brinkmann (2005):

An important feature of human kinds is that they can exert effects on themselves (Martin & Sugarman, 2001). They are affected by their classifications and interact with their classifications, sometimes affecting the classifications themselves. ... This is the looping effect of human kind: 'People classified in a certain way tend to conform to or grow into the ways they are described; but they also evolve in their own ways, so that the classifications and descriptions have to be constantly revised' (Hacking, 1995a, p. 21). And further: 'Inventing or molding a new kind, a new classification, of people or of behavior may create new ways to be a person, new choices to make, for good or evil. There are new descriptions, and hence new actions under a description. (p. 239)' ... As Hacking (1995a) says: 'When new descriptions become available, when they come into circulation, or even when they become the sorts of things that it is all right to say, to think, then there are new things to choose to do. When new intentions

¹² We do not, of course, solely allude to neuroscience here.

¹³ Although we did not find any studies that clearly demonstrate the correlation between neuroscience's view of human mentality and the public's diminished belief in, for instance, self-control, a meta-analysis by Twenge, Zhang and Im (2004) does reveal that people's locus of self-control shifted significantly during the period of 1960–2002, from an internal locus (belief that we are in control of our lives) to an external locus (belief that things besides ourselves control our lives). Whether this has anything to do with the growing acceptance of and trust in neuroscience's view of human nature remains an open question.

¹⁴ We say co-constitute since we are not espousing the representationalist view of the mind. For a critique of representationalism from the enactivist perspective — where cognition (the mind) is seen as co-constituted by the relation between embodied, experiencing agents and their environments — see Varela et al. (1991) and Strel (2016, 2017).

¹⁵ For a similar distinction, see also Fodor (1974).

become open to me, because new descriptions, new concepts, become available to me, I live in a world of new opportunities.’ (p. 236). (pp. 773–774)

A similar point is provocatively put forth by Alasdair MacIntyre: “It makes no difference to water what we call it, but it does make a difference to human kinds. After all, ... molecules don’t read chemistry textbooks, whereas humans do read psychology books that affect their self-understandings” (as cited in Brinkmann, 2005, p. 775); and, we should add, their moral actions.

4. In conclusion: the precarious engagement of neuroscience and the law

All in all, the growing trend of marrying neuroscience and the law has been on the way to creating looping effects in neurolaw. Specifically, the more neuroscience endorses and promotes deterministic, mechanistic, and reductionistic conceptions of the human mind, and the more the public and legal decision-makers accept such claims — and thus (through that) neuroscience is being integrated into the law more and more systematically — the more peoples’ mind, brain, and actions mirror such conceptions of human nature. Further, as this process *loopingly* progresses, the more it would *seem* that neuroscience is correct in claiming that we have and behave as having less agency than we in fact do have (or are able to have), thus making us less responsible for our actions than we are (or could be). Looping effects are probably inevitable in the context of the law; but, of these we should be aware, since the law should remain a contract based on values we *consciously decide* (or agree upon) to preserve.

If we agree with the claim that neuroscience is about human and not natural kinds, and thus cannot be claimed to be discovering objective facts about human mentality and revealing the true causes of human behaviour, we cannot (and should not) reasonably claim that findings about the purportedly objective facts about the brain (extrapolated to the human mind) can be the main or sole reason for changing legal systems, or that they can be accepted readily as evidence of criminal mentality and actions.

Overall, we should be much more cautious about what kind of marriage between neuroscience and the law we would like to endorse, since neuroscience and its findings about (conscious) mental phenomena are not free of values, beliefs, and actions which we apply to them.

Acknowledgements

The work of Olga Markič has been supported by the Slovenian Research Agency (research core funding No. P6-0252).

References

- Aharoni E, Vincent G M, Harenski C L, Calhoun V D, Sinnott-Armstrong W, Gazzaniga M S, Kiehl K A (2013). Neuroprediction of future rearrest. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15): 6223–6228. <http://doi.org/10.1073/pnas.1219302110>
- Bandura A (1989) Human agency in social cognitive theory. *The American Psychologist*, 44(9): 1175–1184. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/2782727>
- Baumeister R F, Masicampo E J, DeWall C N (2009) Prosocial Benefits of Feeling Free: Disbelief in Free Will Increases Aggression and Reduces Helpfulness. *Personality and Social Psychology Bulletin*, 35(2): 260–268. <http://doi.org/10.1177/0146167208327217>
- Bennett M R, Hacker P M S (2003) *Philosophical foundations of neuroscience*. Wiley-Blackwell.
- Brigandt I, Love A (2017) Reductionism in Biology. In ., *The Stanford Encyclopedia of Philosophy*, ed. E N Zalta (Spring 2017 Edition). Retrieved from <https://plato.stanford.edu/archives/spr2017/entries/reduction-biology/>
- Brinkmann S (2014) Natural and Human Kinds. In *Encyclopedia of Critical Psychology*, ed. T Teo, 1220–1225. New York: Springer-Verlag.
- Brinkmann S (2005) Human Kinds and Looping Effects in Psychology. *Theory & Psychology*, 15(6): 769–791. <http://doi.org/10.1177/0959354305059332>
- Catley P (2016) The Future of Neurolaw. *European Journal of Current Legal Issues*, 22(2). Retrieved from <http://webjcli.org/article/view/487/651>
- Catley P, Claydon L (2015). The use of neuroscientific evidence in the courtroom by those accused of criminal offenses in England and Wales. *Journal of Law and the Biosciences*, 2(3): 510–549. <http://doi.org/10.1093/jlb/lsv025>
- Choudhury S, Nagel S K, Slaby J (2009) Critical Neuroscience: Linking Neuroscience and Society through Critical Practice. *BioSocieties*, 4(1): 61–77. <http://doi.org/10.1017/S1745855209006437>

- Church D J (2012) Neuroscience in the Courtroom: An International Concern. *William & Mary Law Review*, 53(5): 1825–1854. Retrieved from <http://scholarship.law.wm.edu/wmlr/vol53/iss5/8>
- Crick F (1994) *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Charles Scribner's Sons.
- Farah M J (2012) Neuroethics: the ethical, legal and societal impact of neuroscience. *Annual Review of Psychology*, 63: 571–591. <https://doi.org/10.1146/annurev.psych.093008.100438>
- Farah M J, Hutchinson J B, Phelps E A, Wagner A D (2014) Functional MRI-based lie detection: scientific and societal challenges. *Nature Reviews Neuroscience*, 15(2): 123–131. <https://doi.org/10.1038/nrn3665>
- Farahany N A (2016) Neuroscience and behavioral genetics in US criminal law: an empirical analysis. *Journal of Law and the Biosciences*, 2(3): 485–509. <http://doi.org/10.1093/jlb/lsv059>
- Farwell L A (2012) Brain fingerprinting: a comprehensive tutorial review of detection of concealed information with event-related brain potentials. *Cognitive Neurodynamics*, 6(2): 115–154. <http://doi.org/10.1007/s11571-012-9192-2>
- Fodor J A (1974) Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese*, 28(2): 97–115.
- Gazzaniga M S, Keil F C, Goodstein J, Rawson E, Gray J R, Passingham R E (2008) The Law and Neuroscience. *Neuron*, 60(3): 412–415. <http://doi.org/10.1016/j.neuron.2008.10.022>
- Glannon W (2009) Our Brains Are Not Us. *Bioethics*, 23(6): 321–329. <http://doi.org/10.1111/j.1467-8519.2009.01727.x>
- Greene J, Cohen J (2004) For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 359(1451): 1775–1785. <http://doi.org/10.1098/rstb.2004.1546>
- Gurley J R, Marcus D K (2008) The effects of neuroimaging and brain injury on insanity defenses. *Behavioral Sciences & the Law*, 26(1): 85–97. <http://doi.org/10.1002/bsl.797>
- Hacking I (1995) The looping effect of human kinds. In *Causal cognition: A multidisciplinary debate*, eds D Sperber, D Premack, A J Premack, 351–383 and 384–394. New York: Oxford University Press.
- Job V, Dweck C S, Walton G M (2010) Ego Depletion—Is It All in Your Head? *Psychological Science*, 21(11): 1686–1693. <http://doi.org/10.1177/0956797610384745>
- Jones O D, Marois R, Farah M J, Greely H T (2013) Law and Neuroscience. *Journal of Neuroscience*, 33(45): 17624–17630. <https://doi.org/10.1523/JNEUROSCI.3254-13.2013>
- Jones O D, Wagner A D, Faigman D L, Raichle M E (2013) Neuroscientists in court. *Nature Reviews Neuroscience*, 14: 730–736. <https://doi.org/10.1038/nrn3585>
- Jones O D, Bonnie R J, Casey B J, Davis A, Faigman D L, Hoffman M, ... Yaffe G (2014) Law and neuroscience: recommendations submitted to the President's Bioethics Commission. *Journal of Law and the Biosciences*, 1(2): 224–236. <http://doi.org/10.1093/jlb/lsv012>
- Kordeš U (2016) Going Beyond Theory: Constructivism and Empirical Phenomenology. *Constructivist Foundations*, 11(2): 375–385. Retrieved from <http://constructivist.info/11/2/375>
- Markič O (2013). The philosophical framework for understanding neuroscientific research. *Interdisciplinary description of complex systems*, 11(4): 351–362. <https://doi.org/10.7906/index.11.4.1>.
- Miller G (2010, December 14) Brain Exam May Have Swayed Jury in Sentencing Convicted Murderer. *Science Magazine*. Retrieved from <http://www.sciencemag.org/news/2010/12/brain-exam-may-have-swayed-jury-sentencing-convicted-murderer>
- Monterosso J, Royzman E B, Schwartz B (2005) Explaining Away Responsibility: Effects of Scientific Explanation on Perceived Culpability. *Ethics & Behavior*, 15(2): 139–158. http://doi.org/10.1207/s15327019eb1502_4
- Morse S J (2015) Criminal Law and Common Sense: An Essay on the Perils and Promise of Neuroscience. *Marquette Law Review*, 99: 39–72. Retrieved from http://scholarship.law.upenn.edu/faculty_scholarship/1609
- Morse S J (2006) Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note. *Ohio State Journal of Criminal Law*, 3: 397–412. Retrieved from http://repository.upenn.edu/neuroethics_pubs/28
- Nahmias E, Coates D J, Kvaran T (2007) Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions. *Midwest Studies in Philosophy*, 31(1): 214–242. <https://doi.org/10.1111/j.1475-4975.2007.00158.x>
- Rigoni D, Kühn S, Gaudino G, Sartori G, Brass M (2012) Reducing self-control by weakening belief

in free will. *Consciousness and Cognition*, 21(3): 1482–1490.

<http://doi.org/10.1016/j.concog.2012.04.004>

Rigoni D, Kühn S, Sartori G, Brass M (2011) Inducing Disbelief in Free Will Alters Brain Correlates of Preconscious Motor Preparation. *Psychological Science*, 22(5): 613–618.

<https://doi.org/10.1177/0956797611405680>

Savani K, Job V (2017) Reverse Ego-Depletion: Acts of Self-Control Can Improve Subsequent Performance in Indian Cultural Contexts. *Journal of Personality and Social Psychology*. Advance online publication. <http://doi.org/10.1037/pspi0000099>

Shariff A F, Greene J D, Karremans J C, Luguri J B, Clark C J, Schooler J W, ... Vohs K D (2014) Free Will and Punishment. *Psychological Science*, 25(8): 1563–1570.

<http://doi.org/10.1177/0956797614534693>

Strle T (2017) On embodying decision-making and the endless circularity of understanding the mind. *Constructivist Foundations*, 13(1): 76–78. Retrieved from <http://constructivist.info/13/1/076>

Strle T (2016) Embodied, enacted and experienced decision-making. *Phainomena*, 25(98-99): 83–107. Retrieved from <http://cepa.info/4038>

Strle T (2013) Why should we study experience more systematically: neurophenomenology and modern cognitive science. *Interdisciplinary Description of Complex systems*, 11(4): 376–390.

<https://doi.org/10.7906/indec.11.4.3>

Twenge J M, Zhang L, Im C (2004) It's Beyond My Control: A Cross-Temporal Meta-Analysis of Increasing Externality in Locus of Control, 1960–2002. *Personality and Social Psychology Review*, 8(3): 308–319.

http://doi.org/10.1207/s15327957pspr0803_5

Varela F J, Thompson E, Rosch E (1991) *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.

Uršič M, Markič O, Ule A (2012). *Mind in nature: from science to philosophy*. New York: Nova Science Publishers.

Vohs K D, Schooler J W (2008) The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating. *Psychological Science*, 19(1): 49–54. <https://doi.org/10.1111/j.1467-9280.2008.02045.x>

Vörös S, Markič O (2014) Photograph of the Soul: Towards a Critical Neuroscience. *Teorija in praksa*, 51(5): 811–831.

PSYCHIATRY AND NEUROLAW AN ESSAY ON THE MIND-BRAIN PROBLEM AND LEGAL PROOF

Drozdstoy St. Stoyanov

MD, PhD, PgCert, IDFAPA

Department of Psychiatry and Medical Psychology

Medical University,

4002 Plovdiv, Bulgaria

e-mail: stojanovpisevski@gmail.com

Abstract:

The aim of this paper is to highlight the rationale behind the use of data from neuroscience, particularly neuroimaging, in psychiatric legal expert procedures and their interference with the mind-brain problem.

The critical argument is that the employment of mental health evaluation of the defendants and/or witnesses as collected with clinical assessment methods in court proceedings should not be considered irrespective to the data from neuroscience. Essentially, neuroscience methods belong to the domain of nomothetic (natural explanatory) knowledge, whereas clinical evaluation methods in psychiatry belong to the domain of intra- and inter-subjective narratives. There exists an explanatory gap between those two groups of disciplines which concerns the ability to translate and integrate data across diverse methodological and terminological systems. Furthermore, it depends largely on the implicit positions in the mind-brain debate and the brain-to-behavior connections, which reflect on the professional and legal reasoning in terms of prioritizing certain solutions or approaches over another in the expert judgements. There are described those tacit positions adopted in the mind-brain debate by different traditions in psychiatry, with special emphasis on reductive and non-reductive forms of physicalism.

In conclusion, a cognitive pluralist stance is adopted which sets priority for the supervenience theory of mind.

Key words: neuroscience, neuroimaging, data translation, supervenience.

1. Introduction. Psychiatry and moral responsibility. Legal arguments for aspirations at more robust normative criteria

Modern psychiatry is in both a crisis of identity and a crisis of confidence (Kendler and Parnas 2015, Stoyanov and Di Nicola, 2017, Di Nicola and Stoyanov, forthcoming). The crisis of confidence in our view refers to limited consistency and coherence of the methods for clinical evaluation as well as classifications and nomenclature in psychiatry. Instead, there are produced temporary conventions of agreement among professional bodies, which are subjected to strong critical consideration (Zachar, Stoyanov, Aragona & Jablensky, 2014). Such limitations create a caveat of doubt in the soundness of forensic psychiatric expertise, which goes far beyond typical forensic applications (e.g. criminal responsibility) to a broader area of moral and legal responsibility entailed from psychiatric judgements.

Normativity in the medical sciences is taken, by implication, as a deviation from the range of body functions, as defined by normal physiology. Similar attempts to define disease in psychiatry remain as yet far out of reach. It is undermined by the (i) lack of a unitary theory of normal psychology and (ii) the dependence of clinical judgement on values.

Psychiatry bears the burden of legal responsibility entailed by the so called “*harm criteria*” of mental disorder, as discussed elsewhere (Wakefield, 1992; Bolton and Hill, 2008; Telles-Correia, 2018), which is apparently unique and inherent to the field of psychopathology. It is not present, or at least not explicit in the traditional sense, in other medical

disciplines. As it is defined, harm may be directed toward the person, or against other people or societal norms. In addition, harm is typically assumed as the direct biological effect of the disease mechanisms rather than as a conscious decision, subjective experience and behavior. We are confronted here with two dimensions of harmful dysfunction which set the background for defining mental disorder.

One is *subjective suffering*, which may be reported by the patient even in the absence of substantial cause located in the space of ‘common sense’. The other is the *potential or actual damage of inter-subjective relationships*, including immediate violation of legal norms and/or psychotically motivated infliction of physical harm on others. What is fundamental for the understanding of both dimensions is their dependence on values. Values are reflective of a variety of social, cultural, religious, educational, etc. needs and preferences, and impact clinical judgement. These values represent “qualia,” or irreducible intra- and inter-subjective entities, which do not have correlates in the natural world. Precisely, these values — irrelevant to biological signatures of disease — motivate the patients or their group to seek mental health service. If the sole approach adopted, a values based assessment would contribute to a more humanistic and empathetic understanding of suffering. On the other hand, it opens room for an indigestible debate on the responsibility for the consequences of human behavior: if values, and only values, constitute clinical judgement, then almost every human action might be understood from some perspective in the particular context of social narratives (Gurova, 2013). This is a fragile topic when it comes to legal and economic responsibilities. Legal judgements delineate such crucial notions as criminal responsibility, social incapacity and disability, and economic judgements, all of which are relevant to most psychopharmacological and psychotherapeutic treatment methods — it is economical evidence about efficiency and efficacy which may or may not justify public mental health policy. Those problems are embodied in an ongoing debate on the implications of neuroscience in mental health judgement (Bickle and Hardcastle, 2012). The main issue is the way in which evidence from neuroscience may be incorporated as the basis for judicial statements and inference in court proceedings, as well as its implementation in legislative acts. It appears to be very complex to bring together subjective (self) reports and legal responsibility, as any form of subjectivity is *par excellence* liable to abuse.

2. Assessment methods in psychiatry

Some of the arguments above, as well as a hopeless effort to elaborate more “scientific”-like methodologies, underpinned an almost 100 years endeavor to decompose narratives into structured quantitative evaluation methods.

Clinical assessment methods in psychiatry are essentially based on the fragmentation of narratives which are derived from professional (observer based interviews), or either informants’ (such as family and community members) or self-evaluating (patients’) reports. Those might be qualitative (phenomenological) or quantitative. The initial material to inform the construction of the clinical tools was in fact comprised of qualitative data: basically, patient’s reports on diagnostic interviews from 1930s, taken in a psychodynamic manner. Those reports, and actually the most common responses, were then “transcribed” into structured self-assessment scales or inventories, e.g. MMPI (1937), BDI (1961), Zung (1965), Von Zerssen (1970), and others. Similar to the ‘inventories’ employed in clinical psychology assessment, psychiatry has developed interviews which basically represent observer-based rating scales.

What happened over the next 40 years is that the validation approach in mental health disciplines is following a pattern “*from scale – into scale*”. For instance, the most popular Montgomery-Asberg depression rating scale (an observer based instrument) is typically validated against clinical self-evaluation tools such as the Beck Depression Inventory and *vice versa*. There is also employed the so called “*informant*” data in the case formulation which is presumed to be objective; however, it also represents certain personal narrative structures, only from a different perspective.

To put it in other words, this means that each clinical tool (both interviews and self-assessment tests) are validated against similar measures, composed of statements and questions derived in various ways from primary sources (Stoyanov, Machamer, & Schaffner, 2012). As it has been highlighted there, the original or primary source of clinical information lies within the domain of narrative. That narrative is fragmented further in order to compose items and arrange them into statistically susceptible units of analysis. Yet, the diagnosis on all levels — from clinical symptoms and signs to nosology — remains a matter of subjective judgment, placed in the “space of reasons.” Further, Broome and Bortolotti (2009) specifically emphasize that “the main object of study

needs to be the person. The normal and the abnormal themselves are normatively defined and are not properties of the brain" (p.370).

In this reference, the authors imply by "*normativity*" social and cultural norms. Indeed, we do not presume that certain behavior or experience is abnormal because of genetic or neuroimaging scans; rather, we presume it based on our inter-subjective background, as located in the domain of social, ethnical, cultural, religious, educational and other values. Biological data in that regard are taken as adjacent and supplementary, and are not considered to have any diagnostic implication. This has been the main reason behind the radical project for Research Domain Criteria (RDoc) of the National Institutes for Mental Health, as proclaimed in 2010 by Cuthbert and Insel. To summarize this framework, it aspires for complete medicalization of psychiatric knowledge by reconstructing its bottom-up causation in a bio-medical manner. It has been assumed by all means as a return to reductive models, such as eliminative materialism and epiphenomenalism, as discussed in the sections to follow.

3. Disciplinary structure of psychiatry. Data translation and reification of diagnosis

Meta-structure is the conceptual organization of a given shared disciplinary field which allows terminology and data transfer across themes of mutual interest. It includes two facets.

One is meta-linguistic compatibility. This refers to the *commensurability* of the notions and methods applied by different composite disciplines. All sub-disciplines which contribute to the explanation of certain process are supposed to have a "manual" for translation in order to safeguard the equal meaning behind the employed special language terms, as well as respective notions. This results in an integrative taxonomy: classification and nomenclature. Psychiatric knowledge is co-produced with neuroscience, the social sciences (e.g. law, economics, sociology), the humanities (literature, art), and the intermediate field of psychology. In order to achieve meta-linguistic compatibility, any shared notion should have similar (if not equal) meaning in each and every of those disciplinary systems. Unfortunately this is not the case. If we take as an example the term "emotion," as well as other relevant notions (e.g. "affect" and "mood"), we may easily discover that there are various and sometimes incompatible definitions for these in any of the disciplinary matrixes

mentioned above. Moreover, there is no shared definition of what should be assumed as a "normal" mood, emotion, or affect. For neuroscience it would be determined by the "firing" of neuronal systems belonging to the Papez circle or the limbic system. Neuroscience acknowledges the fact that such activations occur in the presence of external stimuli which are quite individual; yet, there is a ubiquitous, universal mechanism to process emotions regardless of their valence (it claims), and this mechanism lies inside brain networks. For the social sciences, emotions and affect are quite controversial, and they sometimes have no correspondence whatsoever with the medical or neurobiological definitions. One particular example is the so called "*pathological affect*" which may be observed in perpetrators of crimes, and is taken by the legal disciplines as a formal "excuse" for certain harmful kinds of behavior or as a special psychiatric condition. Yet, pathological affect does not exist as such, neither in psychopathological nomenclature nor in neuroscience, as neither of the two disciplines has identified the range of "*normal*" *affect* as a fundamental prerequisite to define the "pathological dimension".

Second is axiomatic stability, which describes existing consensual causation laws and the criteria for their validity. To put this another way, this delineates the extent to which there is agreement among fields of inquiry (though not among expert bodies, which is more a convention) as to what is the essential mechanism triggering the observed phenomenon. Psychiatry in general is divided by two possible causal theories: one overestimates the agency of mental processes (and respective subjective experiences, conscious or non-conscious, dependent on the theoretical concept), whilst the other regards the neuronal processes as primary causal agent for both normal and abnormal experiences and behavior. This controversy is a specific instantiation of the philosophical mind-brain problem in psychiatry.

With the exception of psycho-organic syndromes, the process of translation and reification in clinical assessment remains as yet far out of reach. All data acquired 'externally' of the domain of clinical psychopathology are regarded as irrelevant to diagnosis, and thus serve for experimental research only. This refers to all recent and current studies in neuroscience which are supposed to be the subject of translation, such as genetics, immunology, electrophysiology (EEG), functional MRI. In reality, however, triangulation is performed within one and the same realm of knowledge, and it is through the

realm of narrative, both professional (interviews) or patient's (self-assessment), whereby these two converge into the quantitative measures named "scales." Scales are usually composed of structured questions/statements (called items) which require different kinds of item response. The dichotomous scale is a two-categories-scale that can be answered by "Yes"/"True" or "No"/"False", but which gives no information about the subjective frequency or degree of the complaints. The Likert scale on the other hand gives the opportunity to express a degree of consent/frequency/degree of the complaints. In this way, the subject is able to give an answer about the frequency of a symptom's manifestation. Usually, the answers on the Likert scale are summed up, and through a standardized coefficient a characteristic's level of manifestation is evaluated.

Although psychological and psychiatric assessment tools (scales) belong nominally to different disciplinary frameworks, they still exert one and the same matrix: the matrix of a statistically quantified narratives. There are poor or no investigations on the role of external validator or validity operation.

If we assume the case of depression, then the translation across disciplines to operationalize diagnostic procedure will appear like this:

- **Psychiatry:** Psychiatry employs a variety of clinical interviews to evaluate depressive disorder. The most common among these are the Montgomery-Asberg Depression Rating Scale (MADRS) and Hamilton scale for depression (HAM-D). These consist of 8-10 items, with the evaluation of each item being applied according to the Likert scale. For instance, a score on MADRS over 20 is considered as a measure of mild or moderate depression, whereas a score above 30 is considered to be an indicator of severe depression.
- **Clinical Psychology:** The usual psychological tests (or inventories) that are implicated in depression are Beck's and Zung's inventories. For the reason of measurement of the outcome of treatment in clinical psychopharmacological trials, there is applied the Depression Scale by Von Zerssen in German speaking countries.

The translational procedure statement here appears like this: *if a depression score on MADRS of 25 is corresponding to a Von Zerssen DS score of 22, it is most likely that the patient has major depressive disorder, current episode of mild depression.*

In this framework there is no agent of reification, and much confusion is caused by diagnostic mistakes and either overestimating or underestimating the clinical data. Any discrepancy/dissociation between the two measurement scores is interpreted as a rule in favor of the rater based evaluation (the interview), though there are no official guidelines to regulate this. As it has already been stated elsewhere, the lack of translation towards other disciplines and reification of psychiatric diagnosis undermines its expert validity in many situations of heavy social, legal, and economic burden, ranging from the choice of drug and therapeutic monitoring to expertise of diagnosis of mental disability and social incapacity.

Of course, there are established post hoc correlations between the different types of clinical tools (interviews and inventories) on one hand and neuroscience on the other; however, they do not converge as diagnosis on the construct of "depression," insofar as the different kinds of measures represent mere statistical correlations with no reference to the mechanism of disorder, and therefore cannot be effectively translated into and embodied in diagnostic standards and procedures. Moreover, there is raised a fundamental issue: what is the *subject of reification that procedures of translation may address*? For example, is it the interview/inventory that is reified by means of functional MRI, or is it *vice versa*? What is the difference between "object" and "subject" of reification? The original assumption would be that the data acquired with a 'narrative' based method comprised of more or less subjective item reports should be the subject, and the object is the data collected with biological, 'nature based' method. Yet, this corresponds to a rather reductionist stance to the problem. The mind-brain dualism promotes another view: there is no room left for any biological reification at all, as the very object of psychiatry happens to be subjective human experience, and it is not liable to any kind of reification and/or validation outside the domain of "narrative."

This leads us to the crucial importance of the mind-brain problem in psychiatry.

3. Mind-brain problem's implications in psychiatric expertise

3.1. Historical roots

We assume that various stances in the mind-brain debate underlie the main diagnostic and therapeutic methods in psychiatry. Whilst biological pharmacotherapy has largely expanded on the basis

of eliminative materialist views, psychotherapy is endorsed on an implicit level by the dualism, and perhaps dual-aspect monism. In fact, *the crisis of confidence* means that we *no longer trust* our clinical evaluation methods, or our therapeutic ones. This undermines our expert statements and their legal authority as well.

In the ages before neuroscience evidence had been delivered — though controversial in the fields of psychology and psychiatry — the mind-brain debate had been transferred from the field of philosophy into the field of psychiatric and neuroscientific explanations. This entails causal structure, mechanisms of production of mental disorders, and thereby clinical decision making based on/informed by different stances in the mind-brain debate.

There are two traditions that may be demarcated in the long and tortuous history of psychiatry with respect to the mind-brain problem (Machamer & Stoyanov, 2009). First, there is the medical tradition, as found in anatomy and physiology. Thomas Willis (1621-1675) is perhaps the best known early modern practitioner of this science. Willis studied the brain most carefully, and in fact compared a normal brain with the abnormalities he found in patients who had congenital mental retardation. His most detailed works on abnormal behaviors are *Pathologiae Cerebri et Nervosi Generis Specimen* (1670) and *Affectionum quae dicuntur hystericæ et hypochondriacæ pathologia spasmodica vindicata, contra responsionem epistolarem Nathanael. Highmori* (1678). Most often, this tradition sees the brain to behavior connection as strictly causal. In one version this tradition is ontologically reductive.

The second tradition relates the brain (and other bodily workings) to the mind, and then the mind to behavior. Perhaps the best known early theorist in this research was René Descartes (1596-1650), who in his *Les Passions de l'Ame* (1649) attempted to describe the bodily bases for human passions, and theorized how unchecked passions lead to abnormal or excessive behaviors. As to how the brain (and body) affects the mind, as well as how the mind then affects behavior, various types of relations are hypothesized in this tradition. In this tradition — known as *Cartesian dualism* — the mind is often treated as a separate ontological kind, and is taken to have representational properties that are responsible for behaviors. In some versions, the way in which the mind brings about behavior is held to be non-causal.

Of course, there are intermediate positions, and some confusing attempts at combination. One such would be the position of *Sigmund Freud* (1856-1939), who held that the mind was explanatorily independent from the brain, but not ontologically. He held that one day we would be able to explain every psychopathology in terms of brain functions, but until that time one needed independent mental constructs to explain the etiology of such pathologies. So Freud was not reifying the mind as a separate ontological entity, but did hold that, due to its representational (or ideational) nature, the mind could be (and for therapeutic purposes had to be) discussed in ways independent of the physiology of the body.

In addition, beyond the frontiers of scientific contexts, there are the spiritual explanations of mental disorders, generated by religious traditions. ‘Treatment’ was by exorcism, though in some forms this spiritual cause may be seen even in the nineteenth century (‘psychics’ and Jacobi in Germany).

During the period of the Enlightenment, a number of attempts were made to bridge brain research to the issues of mental health and disease. Worth mentioning are two of the treatises written by Friedrich von Schiller (1780), which consist of the essentials of later psycho-physiology. Actually, there were no empirical data collected before the beginning of the twentieth century, and the mind-brain relationships were still described and explained through philosophy.

In the eighteenth and nineteenth centuries there was achieved significant progress in the brain-to-brain field of research — with respect to the experimental and descriptive investigation of brain anatomy and physiology themselves. Some of the outstanding contributions were associated with the names of Felix Vicq D’Azyr, F. Majendie, Emile De-Bois Reymond, and others. Nevertheless, most of them were not open to the problems of mental disorders.

3.2. Current implications of the mind-brain debate in psychiatry

One leading tacit position in current psychiatric explanations remains Cartesian dualism. Without explicit commitment or reference to it, most in the psychotherapeutic field are influenced by interactionist dualism. The main postulate of psychotherapy, which is typically taken by implication, is that mind and brain are separate but interacting entities. Psychotherapy claims to address the mind (or psy-

che); however, it admits that the effect from psychotherapeutic interventions is due — at least in part — to modifications of the underlying neural networks, particularly to chemical neuroplasticity. Interestingly, many anti-depressants actually have a similar mechanism of action. Over the past few decades there were developed several trends which aim at experimental confirmation that psychotherapy affects the neural substrate in the first place. Different results have been reported with both psychoanalytic psychotherapy and Cognitive-Behavioral approaches (Morgieva et al., 2014). More specifically, Solms' research (2002) has led into a novel area of trans-disciplinary interactions, called "*Neuro-psychoanalysis*." This triggered an interpretation of the mind-brain physicalism, known as "dual-aspect monism," which assumes the human psyche and brain mechanisms are two manifestations of one and the same core substrate, and it might be regarded as close to Donald Davidson's "anomalous monism." Neuro-psychoanalysis claims to be focused mainly on Sigmund Freud's so called "early" project for "scientific psychology." It is part of the correspondence between Freud and Wilhelm Fließ in the period from 1895 to 1930, and was published in a separate edition in 1950. The "Project" is basically assumed as a variation of the reductive theories of mind and belongs to the classical heritage of psychoanalysis. It has been underestimated by the different psychoanalytic traditions and generations for decades since Freud's death. We owe Karl Pribram (1976) for revisiting this project in the context of scientific progress. Outlined through the social and historical dynamics of the psychoanalytic movement in the twentieth century, the rejection of this fundamental construct in Freudian theory is scrutinized elsewhere.

Instead of a conclusion, we propose a nonclassical interpretation of the actual motives for the revival of the classical understanding of the Freudian conceptualizations at the end of the twentieth century, as embodied in the postmodern discipline called neuro-psychoanalysis (Stoyanov, 2009). This view is in a way controversial since it resembles the outdated concept of "*epi-phenomenalism*," in which human mental life was regarded as collateral by-product of brain activity, in the same way as the bile is by-product of the liver activity (this metaphor is usually ascribed to the neo-Cartesian philosopher Pierre Cabanis). It is also in line with Patricia Churchland's eliminative materialism, although missing their original computational perspective. In

that regard, radical forms of reductive physicalism remain major temptations for modern psychiatry as a way to achieve the coherence and predictable structure of the medical disciplines.

To summarize, mainstream psychotherapy is still committed to one or another extent to a dualistic understanding of the mind and brain. On the other hand, biological psychiatry (often referred to as neuropsychiatry) is looking into various forms of monistic physicalism. The latter might be seen as reductive and non-reductive. Reductive physicalism has been employed in order to provide a sound explanation for mental disorders, and it is presumably an effective approach to manage them (Stoyanov, 2014). In the twentieth century, reductive physicalism developed into several influential traditions relevant to psychiatry: eliminative materialism and the identity theory of mind. Both of these assume the supremacy of neurobiological models, tending to reduce psychological and psychopathological phenomena to events and processes in the brain.

Eliminative materialism is attributed to the works of Patricia Churchland (1989), without particular reference to psychopathology. It comes to a great extent as a logical consequence from the strict Identity Theory of Mind exposed below, with a stronger metaphysical commitment. The main subject was considered to be "*folk psychology*," terms and vocabulary for which were supposed to be dismissed (eliminated) by computational neuroscience. However, as usually happens in science, the core field has been gradually expanded and displaced to occupy new territories of debate. The conceptual background of the Human Brain Project (Markram, 2012) is actually derived from Churchland's theory. Furthermore, most of the pharmacological interventions in psychiatry are also rooted in eliminative materialism on the level of 'tacit knowledge.' Psychopharmacology's modus operandi is developed in statements like "*if we provide a drug to bind with a receptor (sub) population for the substance X, then it will reduce the symptom Y.*" Let us presume that the drug substance is dopamine receptor antagonist, and the symptom in question is verbal hallucinations. In this case, the tacit content of such statement might be reformulated like this: *Verbal hallucinations are caused exclusively by dysfunction/deregulation of the dopamine receptor and its blockade will remove them. Therefore verbal hallucinations are merely and external manifestation of dopamine dysregulation in the brain.*

Essentially, this means that we no longer need to articulate the subjective content (phenomenology) of verbal hallucinations since all of them are completely reducible to biochemical entities. Thus, phenomenological psychopathology, both as method and vocabulary, is useless, and *hearing voices* is nothing else but disturbance of the dopamine metabolism. It is critical to stress here that this inference is taken *without residuum*, i.e. no further social, cultural, or religious interference is assumed as being related on a causal level to production of the phenomenon.

Besides eliminative materialism, there has been introduced another reductive paradigm: identity theory of mind, or ITM (Stoyanov, 2014). ITM has grounded its basic claims in the field of experimental arguments, namely the so called “c-fiber” argument. This argument refers to the differential activation of the c- and a-delta-fibers of the pain perception system. The c-fibers provide slower transduction of the neural signal from the peripheral pain receptors to central areas as compared to a-fibers. Identity theorists claim that the firing of a-delta fibers is the correlate of the physical perception of pain (triggering the adaptive response of withdrawal), whilst ‘firing’ of c-fibers is the immediate correlate of the emotional experience of pain. Identity theory expands this explanation to the entire realm of subjectivity, or consciousness. There are two versions of ITM. The more radical version is the type-type identity as formulated by U.T. Place; the other is token-ITM, as formulated by J.J.C. Smart. Type identity takes for granted that human consciousness is *literally* reducible to lower entities, or levels of explanation, in the same way as lightning might be reduced to ‘electrical discharge’ and heat to ‘molecular motion.’ As it has been stated in the seminal papers of Ullin T. Place (Graham and Valentine, 2004), identity might be regarded as “*compositional*,” where components of lower level phenomena are incorporated into the hierarchy of higher order ones without explicit ontological claims regarding elimination of higher order phenomena (i.e. human consciousness). This is essentially different from *De re* identity where the higher order phenomenon is overruled: mental states and processes are taken to be presumably *equal* to brain ones. This view overlaps with the eliminative materialism as described above.

On the other hand, token identity assumes identity between mental and brain states/processes in a rather *metaphorical* manner. From this perspective,

“brain process” and “sensations” (or emotional states) are understood as two descriptions of one and the same underlying substance in the same way as “Morning star” and “Evening star” both refer to the planet Venus in astronomy. This distinction, however, is not of any particular significance when extrapolated to psychiatry. What actually matters to psychiatry (and to psychology, as well) is what kind of objections might be delivered contra reductive physicalism in general and ITM in particular. There are two main objections of relevance to mental disorders: the presence of qualia and multiple realizability.

Qualia is an argument which is pertinent to phenomenological psychopathology and refers to the qualitative dimension of human experience. For instance, if we accept it provisionally that all verbal acoustic hallucinations (VAH) are produced by one and the same brain mechanism (which is in itself a controversial statement), then what might be the feasible explanation for why they differ so much in terms of their content on the inter-individual level? Some people hear voices of their deceased relatives, others hear the voice of God, or demons etc. Also, some of the patients’ voices are experienced as external to their subjective body construction, as coming from the objective world, while other patients report the voices to be experiences “inside” their head, delivered to or embodied in their subjectivity by means of special devices or telepathy. Through the example of his own psychopathological experiences, this complex phenomenon was described by the Russian psychiatrist Viktor Kandinsky under the term “*pseudo-hallucinations*,” later, it was independently described by Gaetan De Clerambault.

Multiple realizability is an argument derived from neuroscience which refers to the observed fact different loci in the brain are involved in the production of certain psychopathological phenomena (like VAH); also, the data reported in the current literature as correlating to a neural substrate of VAH are highly inconsistent. However, the opposite observation is also true: different forms of VAH psychopathological manifestations (like delusions) converge on the same neural substrate as VAH.

Advances in computational neuroscience, inspired by eliminative materialism have facilitated ground-breaking research programs with far more sophisticated, high-technological data processing facilities, such as the various platforms of Human Brain Project, which are practically designed to accumulate data from many basic disciplines, by all

means, in order to reconstruct a computational, *bottom-up* model of brain functions and disorders which includes mental disorders.

If we step aside from the arena of speculation, any of the above stance in the mind-brain debate is counter-productive in itself, as it reproduces reality in a more or less deterministic, paradigmatic, and ontologically fragile (if not vacuous) manner. This observation is unfortunately valid through all other fields and specialties of medicine, and leads to fragmentation of knowledge and a superficial understanding of human nature in health and disease. Additionally, the diverse implicit positions on the mind-brain debate obstruct the establishment of uniform criteria for validity and validation across disciplines comprising psychiatry.

It seems to be that revisiting an identity theory of mind might be more useful. Actually, there has been formulated a moderate form of physicalism in response to the critical accounts of epiphenomenalism and eliminative materialism that is based on the argument of multiple realizability. This is the anomalous monism and subsequent supervenience theory of mind (STM), as brought forward by Donald Davidson (1974) and defined by Jaegwon Kim (1993, 2010).

Supervenience theory adopts the thesis that any difference in physical properties (neural events or mechanisms) should cause differences in the mental world (states or traits dependent), but not *vice versa*. This means that any contrast measured by means of functional MRI between patients and healthy control groups should be associated with respective differences in the clinical manifestations (observations) in the score on various rating scales. However, not every difference in the clinical observation might be reduced to physical properties of the brain, i.e. if two different clinicians report different scores on Montgomery-Asberg Depression Rating Scale (MADRS) in one and the same patient, this does not necessarily imply contrasts in the BOLD signals on fMRI in the same patients.

There are considered two forms of supervenience: global and local. Transferred to the field of psychiatry and psychology global supervenience may appear like this: *There are differences in patterns of brain activations in all patients with depression when compared to patients with schizophrenia or healthy controls*. This position appears to be false for many empirical and meta-empirical reasons as discussed elsewhere, so we should rather adhere to the local supervenience. In our case it applies like

this: *There are differences in the patterns of brain activations between the item responses on depression inventory as compared to diagnostically neutral items and to paranoid items in patients with presumed depressive disorder when compared to patients with schizophrenia and to healthy controls*.

Contrasted activations in the middle frontal gyrus (MFG) between patients and healthy control groups are immediately associated with clinically significant differences in the score on the Depression Scale (DS) of Von Zerssen, and any difference in the von Zerssen DS is associated with contrast measured in the underlying brain processes with fMRI (Stoyanov et al., 2017; Stoyanov, Kandilarova et al., 2017). The first condition needing to be met in order to maintain the identity thesis here is that the determined differences in the mental states (normal and depressed), as well as the differences in the correlated brain activations, should present not only at the level of statistical aggregates but *on an individual level without any inter-individual variability* (which is certainly not the case). Another condition would be to further investigate necessary and replicable differences in comparison to other constructs employed in psychopathology, e.g. paranoia, anxiety, and others. Our current investigations towards establishment of discriminative validity against the paranoid scale by Von Zerssen are still inconsistent to draw inferences, with interesting contrasts described on single subject level in the hippocampus and amygdale though.

My basic claim here is that, insofar as *local* supervenience theory is an appropriate approach to intra-subjective (*introspective*) evaluation by self-evaluation instruments when combined with fMRI measures as proxy variables, it might be partly liable to empirical reduction according to the *strict identity theory of mind*. The inter-subjective evaluation (observer based interviews) can only supervene in terms of global supervenience on physical (neurobiological) properties. This is entailed from the complex irreducibility of the *reflexive and intuitive space of inter-subjectivity*.

Conclusions and future guidelines:

According to our analysis there are at least two divergent approaches in the context of legal proceedings dependent on the tacit positions adopted by the relevant experts regarding the mind-brain problem.

One would be the position of reductive physicalism (either in its epiphenomenalist or identity theory interpretation). This approach would disre-

gard any differences or variations at the level of subjective reports as embodied in inventories and interviews, and build its judgement solely on findings from neuroimaging and other methods available from modern neuroscientific inquiry. No matter what the total score on clinical evaluation scale might be, the expert will demand results from neurobiological measures to distinguish the healthy person from mentally ill, and to thereby ground his statement on legal responsibility. Such would be in line with Thomas Willis' predictions which essentially converge on the assumption that there are significant organic differences in the structure (macro- or microscopic) of normal and abnormal "brains."

The other would be the tacit dualist position, where the subjective evaluation reports would have equal weight when compared to neuroimaging findings and might be interpreted irrespective of the latter. In contemporary variations of interactionist dualism, a neurobiological correlate would be expected to be found as a substrate of genuinely organic psychiatric conditions such as dementia or delirium, and all others would be regarded as genuinely "mental" conditions without any physical agency present in their causation.

The third view — for which I am sympathetic — would be supervenience: in its local version when it comes to self-evaluation tools, and its global version when it comes to observation based rating scales. Any diagnostically specific score on self-evaluation items (scale) which is considered to be "abnormal" or "pathological" should be underpinned with particular contrasts on blood-oxygenation-level-dependent (BOLD) signals in fMRI performed simultaneously with the same items/scale in patients with different diagnoses. On the other hand, generally defined as "pathological" results from observational interviews, should correlate to a broader basis with anomalies in brain function recorded separately from the interview when compared to healthy population. In both cases we expect that the subjective evaluation should correspond to neuroimaging measures; however, that correspondence might be seen as relatively stronger for self-assessment in comparison with observational methods.

References

Bickle J, Hardcastle V G (2012). *Philosophy of neuroscience*. New York, NY: John Wiley & Sons, Ltd.

Broome M R, Bortolotti L (2009) *Psychiatry as cognitive neuroscience: Philosophical perspectives*. Oxford, UK: Oxford University Press.

Bolton D (2008) *What is mental disorder?: An essay in philosophy, science, and values*. Oxford, UK: Oxford University Press.

Davidson D (1974) Psychology as philosophy. In: *Philosophy of psychology*. London, UK: Palgrave Macmillan: 41–52

Graham G, Valentine E R (eds) (2004) *Identifying the Mind: selected papers of UT Place*. Oxford UK: Oxford University Press.

Churchland P S (1989) *Neurophilosophy: Toward a unified science of the mind-brain*. Cambridge, MA: Massachusetts Institute of Technology Press.

Cuthbert B N, Insel T R (2013) Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC medicine*, 11(1): 126.

Di Nicola V, Stoyanov D (2018) *Psychiatry in Crisis: At the Crossroads of Neuroscience, Social Sciences and the Humanities*. Springer International Publishing, forthcoming

Gurova L (2013) Understanding it makes it normal: is it a reasoning fallacy or not?. *Journal of evaluation in clinical practice*, 19(3): 524–527.

Kendler K S, Parnas J (eds) (2015) *Philosophical issues in psychiatry: Explanation, phenomenology, and nosology*. Baltimore, MD: Johns Hopkins University Press.

Kim J (2010) *Essays in the metaphysics of mind*. New York, NY: Oxford University Press.

Kim J (1993). *Supervenience and mind: Selected philosophical essays*. New York, NY: Cambridge University Press.

Machamer P, D S Stoyanov (2009) The scientific project of psychiatry, *Asklepios*, III(XXII): 51–56.

Markram H (2012) The human brain project. *Scientific American*, 306(6): 50–55.

Morgiève M, et al. (2014) Dynamics of psychotherapy-related cerebral haemodynamic changes in obsessive compulsive disorder using a personalized exposure task in functional magnetic resonance imaging. *Psychological medicine*, 44(7): 1461–1473.

Pribram K H, Gill M M (1976) *Freud's "Project" re-assessed: preface to contemporary cognitive theory and neuropsychology*. New York, NY: Basic Books.

Schiller J C F (1959) *Versuch über den Zusammenhang der thierschen Natur des Menschen mit seiner geistigen*. Ingelheim: Boehringer.

Solms M, Turnbull O (2002) *The brain and the inner world: an introduction to the neuroscience of subjective experience*. New York: Other Press.

Stoyanov D (2009) Why we are revisiting the classical texts of psychoanalysis? *Sociological Problems*, 1–2(41): 170–181.

Stoyanov D (2014) A linkage of mind and brain: towards translational validity between neurobiology and psychiatry. *Biomedical Reviews*, 22: 65–76.

Stoyanov D, et al. (2017) Towards translational cross-validation of clinical psychological tests and fMRI: Experimental implementation. *Comptes rendus de L'Academie bulgare des Sciences*, 70(6): 879–884.

Stoyanov D, Kandilarova S, Borgwardt S (2017) Translational functional neuroimaging in the explanation of depression. *Balkan Medical Journal*, 34(6): 493–503

Stoyanov D, Machamer P K, Schaffner K F (2012) Rendering clinical psychology an evidence-based scientific discipline: A case study. *Journal of Evaluation in Clinical Practice*, 18: 149–154.

Stoyanov D S, V Di Nicola (2017) Psychiatry in crisis: Epistemological and ontological concerns. *American Journal of Psychiatry and Neuroscience*. Special Issue: “In and out of Your Mind” Abstracts of 1st Eastern European Conference of Mental Health, 5(6–1): 6.

Telles-Correia D (2018) Mental disorder: Are we moving away from distress and disability? *J Eval Clin Pract*. Online First, 1–5
<https://doi.org/10.1111/jep.12871>

von Zerssen D (1986) Clinical self-rating scales (CSRS) of the Munich psychiatric information system (PSYCHIS München). In *Assessment of depression*, 270–303. Berlin: Heidelberg, Springer.

Wakefield J C (1992) The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist*, 47(3): 373.

Zachar P, Stoyanov D S, Aragona M, Jablensky A (eds) (2014) *Alternative perspectives on psychiatric validation: DSM, ICD, RDoC, and Beyond*. Oxford, Oxford University Press.

Zung W W (1965) A self-rating depression scale. *Archives of general psychiatry*, 12(1): 63–70.

NATURALIZING THE SUBJECTIVE SIDE OF THE CRIME: A FEW INTRODUCTORY REMARKS ON THE ROLE OF CONSCIOUSNESS IN CRIMINAL LAW BASED ON AMERICAN AND POLISH EXAMPLES

Bartosz Janik, Maciej Próchnicki

Jagiellonian University, Kraków, Poland*

e-mails: bartosz.janik@doctoral.uj.edu.pl, maciej.m.prochnicki@gmail.com

Abstract:

Advancements in neuroscience cast new light on the functioning of the human mind. This is especially important within the context of criminal law, wherein consciousness plays a crucial role in determining criminal responsibility. Yet, there are some caveats in the direct application of these new findings, most of which are related to the specific conceptual framework of law based upon commonsense knowledge and (sometimes) outdated psychology. This framework has also produced different doctrines of interpretation in the systems of common and civil law. Moreover, the goals of the law are to some extent different from scientific research on the brain. The aim of this study is to assess to what extent and under what interpretation scientific knowledge concerning consciousness might be useful for legal purposes, especially for the criminal law. Our assessment is that most of the current concepts of criminal law are directly related to outdated psychological and neuroscientific theories, and that the content of those concepts should be updated according to the newest scientific findings while remaining in accordance with the primary functions of criminal law.

Key words: neurolaw, consciousness, criminal law, comparative law.

1. Introduction

Achievements in modern neuroscience have shed new light on the functioning of the human mind. Nevertheless, there are some mysteries which remain to be solved. The most complex of these is the phenomenon of consciousness: despite attempts

to find the so-called neural correlates of consciousness (i.e. the activity in the specific parts of the brain that are necessary for awareness or specific perceptual experiences), the functioning of consciousness remains one of the most discussed and controversial topics in cognitive science (Chalmers, 2000).

From the legal perspective, questions about the state of mind of a perpetrator is one of the most important issues when assessing criminal responsibility. Advances in neuroscience can allegedly help us determine whether the defendant had specific knowledge at the time of the crime and whether they could control their behaviour; as such, it is clear that modern science provides us with increasingly accurate models of the nature of the human mental sphere. Criminal law, on the other hand, is often based on old psychological concepts and commonsense terminology. The idea of the naturalization of criminal law — which would aim at fully harmonizing law with the state of neuroscience — is very tempting, albeit it might cause considerable trouble. What is required is the bringing together of three different conceptual schemes: the scientific scheme, which aims to explain and predict phenomena in the natural world; the legal scheme, which aims to deliver justice; and the ordinary scheme, the role of which is to allow people to communicate with each other and to function in the society. The last one manifests itself through so-called “folk psychology”: the universal ability to predict and explain the human behaviour, which often lacks scientific accuracy and has a limited predictive power (Stich & Nichols, 2003).

* The authors are funded from the Foundation for Polish Science, 'Mistrz' program (no. 2/2015): 'Law & Neuroscience: A New Paradigm in the Philosophy of Law'.

Criminal law generally adopts the commonsense understanding of the mind (Sifferd, 2006). The important question is this: should we abandon the commonsense understanding of the mind in the light of new scientific theories? Or, do the specific functions of criminal law practically exclude the idea of filling it with neurobiological terminology? By analyzing the role of the commonsense concept of consciousness, we will try to show how advances in modern cognitive science can aid criminal law.

Before proceeding, we should point to how consciousness is perceived as the *mens rea* within the context of criminal law — or the Anglo-American concept of “guilty mind” — a theme which is also present in other modern civil law systems. Polish and American criminal law systems will be provided as examples for the analysis of the function played by folk-psychological elements of consciousness in legal settings. They will also be employed to provide a background against which the difficulties of incorporating elements of modern neuroscientific findings into criminal law system will be illustrated. Finally, we will present a few ideas pertaining to how these achievements can help in assessing criminal responsibility.

2. The concept of consciousness in criminal law

Consciousness, understood loosely in the terms of folk psychology, is a crucial element to assess whether a crime (which will be understood generally in this article as an act prohibited by criminal law, including misdemeanors) occurred, as well as its gravity. Intuitively, the legal concept of “crime” boils down to two elements: the general cognizance pertaining to a deed committed by a person, and the freedom of action to commit it (Shen et al., 2011). According to commonsense psychology, people act on the basis of their beliefs and desires, which is to say that their actions are targeted towards certain results (i.e. they have intentions). Given this model of human agency — in order to organize a community — specific social norms need to emerge in order to ascertain order and facilitate cooperation.¹ Therefore, we want to ascribe responsibility to people, the role of which is to create a necessary condition of punishment. By creating such a condition, we seek to ensure that our normative system is both just and efficient. Criminal law shaped in this way aims at

pursuing specific goals. There is a variety of functions of criminal law. In the Polish criminal law doctrine, Włodzimierz Wróbel and Andrzej Zoll (2011) describe four main functions of criminal law: protective, equitable, guarantee and compensative.

Rules of criminal law protect certain legal values. In order to do so, a potential wrongdoer is discouraged from violating them by threat of punishment. This is based on an assumption that one makes a decision about one’s own behaviour with its consequences in mind. The protective function of the law is realized by the deterrence, the role of which is twofold: general prevention ought to affect all potential criminals with theoretically inevitable sanctions, and individual prevention is directed at the person who has already committed a crime — it should cause fear of another punishment in the event of committing another crime. As we can see, the preventive role requires awareness of the consequences of an action from the members of a society.

The equitable function of criminal law (i.e. ensuring the social feeling of justice and fair retribution) also requires taking consciousness into consideration: the basic condition of justice would not be realized if a person without explicit knowledge of the meaning of their behaviour were treated the same way as a fully conscious perpetrator.

Also, the guarantee and compensative roles are based on the commonsense understanding of human behaviour. The guarantee function, setting the limits of penal interventions in terms of the principles of a democratic legal state² and human rights, puts the emphasis on the principle of definiteness of criminal statutes — or, that rules of criminal law should be understandable and precise. Once again, it assumes that the addressee of criminal law can acknowledge it and direct their behaviour according to the law. The compensative role, which brings out the importance of recompense, also underlines the element of repentance and forgiveness, both of which require awareness of the significance of action within the physical and social world.

Modern criminal law systems are based on the aforementioned two conditions of culpability, which seem universal and commonsense (Shen et al., 2011). *Mens rea* refers to “guilty mind”, a state of mind covering the meaning of the act (for instance knowledge or intent). *Actus reus* (“guilty act”) describes the act as free-willed (unlike, for example, automatic body movements, reflex actions or con-

¹ For an analysis of evolutionary genesis of legal norms see e.g. Zaluski (2009).

² Or *Rechtsstaat* (in German).

vulsions). In Polish criminal law, the equivalent of *mens rea* is described as the subjective side of a crime — the inner, psychological element: the attitude of the wrongdoer toward her act. The voluntariness of an act is in fact an element of the objective side of a crime, because the freedom of physical action can be turned off only by physical duress (*vis absoluta*) and not by psychological duress (*vis compulsiva*). A person acting under psychological duress (for example, a direct threat) commits an act in the sense of criminal law, but can be exculpated on the other grounds (Wróbel & Zoll, 2011). It is noteworthy that the element of consciousness is not only crucial for the subjective side of crime, but also appears in the element of *actus reus* (Pardo & Patterson, 2015). Voluntariness presupposes the conscious control of one's actions. As noted by Michael Pardo and Dennis Patterson, the minimal conditions to classify an act as willed are twofold: the power to act, as well the power to refrain from acting (Pardo & Patterson, 2015). Even these simple conditions can create some philosophical problems (Frankfurt, 1969). As noted by these authors, advances in neuroscience can be helpful to determine that the wrongdoer lacked the minimal conscious capability to perform voluntary act. The voluntariness of an act may be problematic if we adopt a radical, deterministic view of the world (Aharoni et al., 2008); but, such a case presents potential solutions for saving the traditional concept of free will in criminal law exist. One is compatibilism, i.e. the claim that determinism and free will are not mutually exclusive. The other might be connected with creating specific normative standards (which would in fact be a case of pragmatic compatibilism: to fulfill the goals of criminal law policy, we would delineate a borderline for those cases in which we would hold some people responsible for their actions).

The problems referring to *mens rea* seem more tractable than the whole debate on free will, and the advances in neuroscientific accounts of consciousness are promising in this context. Yet, some questions arise: how can the progress of the cognitive sciences aid criminal law; and, should the cognitive sciences influence the shape of criminal codes or doctrinal views? In order to answer these questions, we need to examine the role of consciousness from the subjective side of a crime.

As remarked by Gideon Yaffe, the American Model Penal Code provides no definition of consciousness. Instead, it provides a list of examples for when a condition of being conscious is not satisfied:

reflex actions, somnambulism, hypnosis, etc. (Yaffe, 2011). It provides four types of culpability: intent, knowledge, recklessness, and negligence (Model Penal Code §2.02). A person acts purposely when it is his or her goal (a conscious object) to cause a specific result or to pursue the nature of a mode of conduct, or they have in mind the attendant consequences (or hope that they exist). The condition of knowledge relates to specific material elements of a crime, which could be either a contemporary circumstance or a practically certain consequence — if one is aware of them, one acts knowingly. Recklessness refers to disregarding some substantial, unjustifiable risk linked to the existence of some material element of a crime (or the appearance of it as a result) in specific circumstances. Negligence has a purely normative character — some person in a specific situation should have known that this risk existed. To some extent, these conditions overlap — to act purposefully, one must know their purpose.

The Polish Penal Code uses similar elements when describing the mental sphere of a perpetrator. A slight but important difference is how the subjective side of a crime is not a form of culpability: guilt in Polish law is defined purely normatively (Lachowski, 2015) — it is the ultimate charge that a model, law-abiding person (a solid professional in a specific domain, cautious and respecting legally relevant values) would not have chosen to engage in a given form of conduct in that specific situation (Wróbel & Zoll, 2011). The Polish Penal Code divides acts into two groups: committed with intent (intentionality) or without it (unintentionality) (Polish Penal Code, Article 9). Intent is similarly bifurcated: intent is defined as having the will to commit an act, or as accepting the possibility of perpetrating an act when it was foreseeable. The first type of intent is described in the literature as direct intent (*dolus directus*), and the second is called an eventual intent (*dolus eventualis*).³ It is noted that, when assessing the intent, we should analyze two aspects: the intellectual — the perpetrator's awareness of the fragment of the world encompassed by the structure of a crime — and the volitional — the person's psychical attitude toward an act. These two

³ Many authors note that the word used to phrase the concept of *dolus eventualis* - "godzenie się" is somewhat vague, especially from the psychological viewpoint. It could be described as a kind of a passive approval towards the hypothetical possibility of committing the act. On the theoretical controversies about this concept see e.g. Dębski (2017).

aspects reflect the classic intuitions of the Humean theory of moral motivation: belief is not sufficient, instead requiring the presence of a separate state of mind to account for motivation (i.e. desire) (Smith, 1987). The abovementioned distinction between awareness and psychical attitude towards an act might be considered in terms of belief about the perceived reality and desire for some state regarding it. It is interesting to note that the Humean theory of motivation gives a simple way to account for the differences between those two mental states by postulating that their difference consists in the direction belief-aims to fit the world and desire-aims to change the world (Anscombe, 1963). Thus, it turns out that the principal structure of intent mirrors the traditional view of a philosophical theory of motivation.

The Polish Penal Code does not use the terms *niedbalstwo* (negligence) and *lekkomyślność* (recklessness),⁴ as they are the terms used in tort law. Instead, it introduces two equivalents: a prohibited act can be committed without intent when a perpetrator does not have the intent to do so, yet does so as a result of not being careful enough in some manner required under the circumstances, though they could (conscious act without intent, the counterpart of recklessness) or should (unconscious act without intent, the counterpart of negligence) have foreseen the possibility of committing the prohibited act.⁵ As we can see, both the Polish and American constructions of *mens rea* introduce some normative criteria (negligence or unconscious unintentionality) aside from factual evaluation of the mental state of the perpetrator. Having analyzed these constructions, we can conclude that the following factors relating to the perpetrator's mental sphere are legally relevant: first, beliefs about attendant circumstances and the significance of the act; second, beliefs about the results of actions; third, the goal that the act ought to

achieve, and finally, the inner evaluation of the act by the perpetrator.

A more problematic issue concerns the overlapping and intermingling of these questions with the legal notion of insanity. Insanity, like the aforementioned constructions of *mens rea*, is a legal term that does not have its direct counterpart in psychology and medicine; yet, the expertise exhibited in these fields is essential to identify this state (Pardo & Patterson, 2015). Insanity describes a general lack of ability to control and understand the meaning of one's behaviour, typically due to some kind or set of psychiatric conditions. Because a person should not be held responsible for his or her actions if such actions were the result of a mental disturbance, this notion is used as an exculpatory condition. Given such a case, he or she should be the subject of the proper medical treatment, rather than some form of punishment.

As pointed out by M. Pardo and D. Patterson, despite the fact that the aforementioned terms relate to the evaluation of mental states, insanity and *mens rea* are different legal institutions, and one of them does not necessarily entail the other (Pardo & Patterson, 2015). Indeed, one can be classified as insane and still qualify under *mens rea* conditions: a mentally ill defendant may still possess a clarity of thought and action regarding a specific act (e.g. in a period of so-called *lucidum intervallum*, or a "lucid interval" in some disease).

The Polish Penal Code states that any prohibited act committed by a person who was not capable of controlling it or recognizing the significance of such an act shall not be qualified as an offence. This rule provides the following reasons for a lack of capability: mental disease, mental deficiency, or other mental disturbance (Polish Penal Code, Article 31). Additionally, this rule has both cognitive (recognizing the significance of an act) and volitional (controlling one's behaviour) elements/aspects, which makes it similar to the regulation of the American Model Penal Code. The other classic standard in common law, the M'Naghten test, emphasizes only the first aspect, i.e. the inability to identify one's behaviour as wrong (Pardo & Patterson, 2015).

Still, there are a number of problems when assessing insanity. Focusing on the actual mental state of a perpetrator during an act supposedly resulting from some psychological disturbance could be difficult from an evidentiary point of view, as well as being potentially too broad: such would exclude cases in which a perpetrator may be held responsible

⁴ These concepts were present in the former criminal code of 1969, but were removed, as the construction of intent was reformulated. The main idea behind this change was that it is generally irrelevant for classifying the subjective side as "unintentionality", whether a person acted negligently or recklessly. See Giezek (1998).

⁵ It is noteworthy, that the category of acts committed without intent is a quite complex construction, rather than being a simple lack of intent (understood in the legal sense). For example, it incorporates, inter alia, some elements of the objective side of the crime - "not being enough careful in the manner required under the circumstances". See e.g. Majewski (2012)

for the circumstances leading to this disturbance. On the other hand, focusing solely on some source of insanity (such as mental illnesses) allows for the justification of even those people whose mental state during the act were not influenced by some illness (Wróbel & Zoll, 2011). Both American and Polish models of insanity combine these elements. As remarked by M. Pardo and D. Patterson, two components of insanity tests need to be fulfilled: a source regarding those psychiatric causes impairing normal functioning, and the capacity which provides the lack of ability to recognize the meaning of an act for the perpetrator (Pardo & Patterson, 2015).

The concepts of *mens rea* and insanity are specific, autonomous legal constructions, filled with obsolete notions taken from nineteenth century psychology, but with no direct reference to contemporary research in psychology, psychiatry, or neuroscience (Denno, 2016). This can lead to several important practical problems. As Deborah Denno points out, the main sin of legal doctrine regarding conscious behaviour is the attachment to binary classifications: voluntary versus involuntary and conscious versus unconscious (Denno, 2016). These dichotomies, unsupported in the light of current research, may cause significant inconsistencies in judicial classifications (Denno, 2003). This could bring about grave consequences, ranging from acquittal to institutionalization in cases of selected mental problems (Denno, 2003). It is true that consciousness is a complex phenomenon, and there are many borderline cases in which it is hard to assess whether the conditions such as knowledge or intent are met. For instance, switching lanes in traffic is often done without reflection, being an instance of semi-automatism; yet, it is definitely a voluntary act in the sense of criminal law (Pardo & Patterson, 2015). This simple example is particularly useful because it can also show individual differences. A professional driver going on a daily routine probably would not even be aware of it, hardly remembering changing the lane; in contrast, a young and inexperienced driver may need to put considerable more conscious psychical effort into such maneuvering. Considering this, Denno proposes two ways in which neuroscientific research should reform law.

Firstly, acknowledging the multi-faceted and gradual (rather than dichotomous) nature of consciousness points toward the introduction of a third category: acts that are semi-voluntary (or semi-conscious) (Denno, 2003). Such a three-tiered approach is a step forward, but it still faces the prob-

lem of borderline cases. Rather than substituting two categories with three, the law should acknowledge that consciousness is a matter of degree. Such a rule is present in the Polish Penal Code regulation concerning insanity: if the perpetrator's capability to either recognize the significance of an act or to control an act is deemed diminished to a significant extent, the court may apply extraordinary mitigation of the punishment (Polish Penal Code, Article 31 §2). The second point made by Denno, answering the question of how neuroscience could help criminal law, is to create a scientific theory of mental states (Denno, 2016). Still, while updating the scientific (i.e. psychological models) of behaviour embedded in legal settings with the newest theories of consciousness may seem a good idea, it encounters a lot of difficulties.

3. Can modern research on consciousness be helpful?

As we can see, some traditional psychological accounts which were the scientific background of legal texts are definitely outdated. In this case, modernizing the shape of criminal law to update its scientific basis may seem like a necessary venture. However, adapting modern neuroscientific research directly may pose a number of problems. Indeed, David Chalmers divided problems relating to the nature of consciousness into two groups, easy and hard (Chalmers, 1995). Easy problems include questions such as: how can we discriminate or categorize environmental stimuli?; what is the nature of attention or control of one's behaviour?; what are the differences between different states of consciousness, e.g. sleep and being awake? The hard problem is the quest for explanation of the phenomenal side of consciousness — the subjective side of an experience (so-called *qualia*). Another interesting distinction in the field of consciousness was made by Ned Block: the phenomenal aspect (*p-consciousness*), describing *qualia*, and the access aspect (*a-consciousness*), referring to the mental function of processing and using pieces of information (Block, 1995).

Fortunately, all the interesting issues from the legal viewpoint fall into the category of easy problems, mostly related to the "access" aspect of consciousness. From the criminal law perspective we just need to know and explain how the function of processing the information about the material elements of a crime works, and in what way it can steer one's behaviour. This is by no means a simple scien-

tific challenge. Easy problems focus on specific mental functions — we have some data and local psychological or neuroscientific theories on them, yet they are mostly far from being solved. Rather than focusing on creating some grand theory on consciousness, criminal law would benefit more from acknowledging up-to-date research on selected “easy problems.” The complexity of the topic of consciousness makes it very hard to create one, unified research paradigm — there are a lot of philosophical, psychological, and neuroscientific theories that try to explain consciousness from different perspectives; but, they remain to a large extent incommensurable (Van Gulick, 2017). The state of the art in this discipline could be described as remaining at the pre-paradigmatic stage, to use Thomas Kuhn’s terms (Kuhn, 2012).

For example, let us take one of the most promising neurobiological theories on consciousness: Giulio Tononi’s integrated information theory (Tononi, 2008). This theory aims to somehow quantify the level of consciousness that would seem ideal for legal purposes by creating normative standards for the minimum level needed to fulfill the criteria for voluntary, semi-voluntary, and non-voluntary acts. Yet, the mathematical goal of the theory seems too abstract for legal purposes — the factor of integrated information tells us almost nothing regarding whether the perpetrator has known certain facts, or even how, which is crucial in the context of criminal law.

Modern neuroscientific proposals of a general theory on consciousness are still in their infancy, with other notable examples including Francis Crick and Christof Koch’s framework based on the analysis of interactions between coalitions of neurons (Crick & Koch, 2003), as well as “consciousness as a workspace” theories, e.g. Bernard Baars’ Global Workspace Theory (2005), as well as Stanislas Dehaene and Lionel Naccache’s framework (2001). However, one might argue that the proper way to understand the problem of consciousness in criminal law is by employing the Higher Order Theories of consciousness. Those theories argue that conscious awareness is a phenomenon that can be explained by using a two-level framework. In essence, conscious mental states are representational states of which subjects are aware (Carruthers, 2007). The criminal law framework mirrors the two level structure of those theories. A detailed interpretation of this issue is beyond the scope of this paper; still, it might be an interesting direction for the future research.

Despite theoretical problems with choosing the best theory of consciousness, the search for the neural correlates of consciousness can still be very helpful for legal purposes on the other grounds. The main role would be to aid the psychiatric and psychological expertise. Most of the helpful research has a neuropsychological character — showing various deficiencies in the perpetrator nervous system. Such evidence may be useful in proving that, at the time of the act, a person was unable to be in a particular mental state, which would be a requisite for holding them responsible for the act (Pardo & Patterson, 2015). Of course, the neural abnormalities need to have enduring characters (for example lesions or brain tumors) in order to do so since the evidence is always presented *ex post*. Determining the mental state at the moment of the crime without disproving the general capacity is, of course, far more problematic (Pardo & Patterson, 2015).

There is also one more line of research that could contribute to evidentiary measures. The field of neuroscience-based lie detection tries to apply technologies of brain scanning (mainly fMRI and EEG) to polygraph testing. For example, so-called *Brain Fingerprinting* is based on an assumption that a specific type of electrical activity in the brain, the P300 wave, which is a sign that a selected stimuli is familiar to the person being tested, could be used as a Guilty Knowledge Test (Farwell, 2012). These new methods are still very controversial and inherit most of the disadvantages related to traditional polygraph testing (Rosenfeld, 2005).

The use of typical “neuroscientific evidence” like brain scans has met with some skeptical voices. One of the threats was called the “christmas tree phenomenon” — the concern that colorful brain images could make a much more powerful impression on judges, and especially juries, than they should due to the shortcomings of brain monitoring techniques. Of course, brain scans can only be used in an auxiliary role in psychiatric or neuropsychological expertise. However, this concern was misplaced, as jurors are far more critical in assessing brain scans than it seemed (Denno, 2016). In a recent empirical study it was shown that the: “mere display of a neuroimage did not increase jurors’ willingness to find a defendant not guilty by reason of insanity.” (Schweitzer & Saks, 2011) This might suggest the conclusion that, despite the fact of how the importance of neuroimaging evidence in forensic expertise grows, jurors still rely mainly on traditional evidence. This might be due to the fact of the overall

unfamiliarity with the neuroscientific research specificity, or even general skepticism about the field (Schweitzer & Saks, 2011).

4. Conclusions

Consciousness plays an imminent role in the assessment of criminal responsibility. Yet, the subjective side of a crime is often based on folk psychology and outdated psychological views in modern legal systems. The newest research in cognitive science may prove to be helpful in harmonizing the understanding of the legal concepts of *mens rea*, as well as insanity and *actus reus*. One of the most important facts is that consciousness is not a binary state — it has a gradual and complex character. Nonetheless, the full naturalization of these legal concepts is pointless. Firstly, with regard to the consciousness research, it is too early to point out a single, unified paradigm that could be adapted as a functional model by the law. Secondly, one might argue that the law needs to use vernacular language, rather than sophisticated scientific terminology, because it has to be understandable. Judges and jurors lack specific knowledge in this field and it is not their task to explain the mental states of the defendant. Moreover, in order to ensure the guaranteed function, criminal law should be especially clear to the addressees of its rules: laypeople, who function in the social order following their folk psychological model of human behaviour. As shown by empirical studies (Shen, 2011), laypeople perform quite well in understanding the current legal intricacies of *mens rea*.

This does not mean that we should not educate judges and jurors. Quite the contrary: the newest achievements in the mind sciences can help in explaining human behaviour, which is crucial in determining criminal responsibility. What is more, judges and jurors need to critically assess opinions presented by expert witnesses — psychiatrists, psychologists, neurologists, etc. The role of expert witnesses cannot be overestimated, especially in hard, borderline cases. Their opinions should nonetheless be reviewed by other specialists, as well as judges, because psychiatric expertise is hardly unanimous (Gowensmith et al., 2013), and the low quality of expertise may have grave consequences (Denno, 2016).

References

- Aharoni E, Funk C, Sinnott-Armstrong W, Gazzaniga M (2008) Can neurological evidence help courts assess criminal responsibility? Lessons from law and neuroscience. *Annals of the New York Academy of Sciences*, 1124(1): 145–160.
- Anscombe G E M (1957) *Intention*. Harvard University Press.
- Baars B J (2005) Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in Brain Research*, 150: 45–53.
- Block N (1995) On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2): 227–247.
- Carruthers P (2007) Higher-order theories of consciousness. *The Blackwell companion to consciousness*, 277–286.
- Chalmers D J (2000) What is a neural correlate of consciousness? *Neural correlates of consciousness: Empirical and conceptual questions*, 17–40.
- Chalmers D J (1995) Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3): 200–219.
- Crick F, Koch C (2003) A framework for consciousness. *Nature Neuroscience*, 6(2): 119.
- Dehaene S, Naccache L (2001) Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79 (1–2): 1–37.
- Denno D W (2002) Crime and consciousness: Science and involuntary acts. *Minnesota Law Review*, 87: 269.
- Denno D W (2003) A mind to blame: New views on involuntary acts. *Behavioral Sciences & the Law*, 21(5): 601–618.
- Denno D W (2016) The place for neuroscience in criminal law. In *Philosophical foundations of law and neuroscience*, eds D Patterson, M S Pardo, 69–83. New York, NY: Oxford University Press.
- Dębski R (2017) Przestępstwo i jego struktura [The crime and its structure]. In *Zasady odpowiedzialności. System Prawa Karnego. Tom 3., Nauka o przestępstwie*, ed. R Dębski, 1–496. Warszawa, PL: C. H. Beck.
- Farwell L A (2012) Brain fingerprinting: a comprehensive tutorial review of detection of concealed information with event-related brain potentials. *Cognitive Neurodynamics*, 6(2): 115–154.
- Frankfurt H G (2013) Alternate possibilities and moral responsibility. *The philosophy of free will*:

Essential readings from the contemporary debates, 139–148.

Gowensmith W N, Murrie D C, Boccaccini M T (2013) How reliable are forensic evaluations of legal sanity? *Law and Human Behavior*, 37(2): 98.

Giezek J (1998) Kilka uwag o ujęciu nieumyślności w nowym kodeksie karnym. [A few remarks on the conceptualization of unintentionality in the new penal code]. In *Nowa Kodyfikacja Prawa Karnego. Tom II*, ed. L Bogunia, 9–20. Wrocław, PL: Wydawnictwo Uniwersytetu Wrocławskiego.

Kuhn T S (2012) *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Lachowski J (2016) Negligence in Polish and English Criminal Law. *Comparative Law Review*, 20: 87–102.

Majewski J (2012) Nieumyślność a brak umyślności. [Unintentionality and the lack of intent] In *Nieumyślność: Pokłosie VIII Bielańskiego Kolokwium Karnistycznego*, ed. J Majewski, 35–47. Toruń, PL: Dom Organizatora.

Model Penal Code and Commentaries (1985) Philadelphia, PA: The American Law Institute.

Mueller P A, Solan L M, Darley J M (2012) When does knowledge become intent? Perceiving the minds of wrongdoers. *Journal of Empirical Legal Studies*, 9(4): 859–892.

Pardo M S, Patterson D (2015) *Minds, brains, and law: The conceptual foundations of law and neuroscience*. New York, NY: Oxford University Press.

Penal Code of the Republic of Poland [Ustawa z dnia 6 czerwca 1997 r. Kodeks karny]. Dz.U.2016.1137.

Rosenfeld J P (2005) Brain fingerprinting: A critical analysis. *The Scientific Review of Mental Health Practice*, 4(1): 20–37.

Schweitzer N J, Saks M J (2011) Neuroimage evidence and the insanity defense. *Behavioral Sciences & the Law*, 29(4): 592–607.

Shen F X, Hoffman M B, Jones O D, Greene J D (2011) Sorting guilty minds. *New York University Law Review*, 86: 1306.

Sifferd K L (2006) In defense of the use of commonsense psychology in the criminal law. *Law and Philosophy*, 25(6): 571–612.

Smith M (1987) The Humean theory of motivation. *Mind*, 96(381): 36–61.

Stich S, Nichols S (2003) Folk psychology. In *The Blackwell guide to philosophy of mind*, eds S Stich, T A Warfield, 235–255. Oxford, UK: Basil Blackwell.

Tononi G (2008) Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin*, 215(3): 216–242.

Van Gulick R (Summer 2017) Consciousness. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from <https://plato.stanford.edu/archives/sum2017/entries/consciousness/>.

Wróbel W, Zoll A (2011) Polskie prawo karne. Część ogólna, [Polish Criminal Law. General Part]. Kraków, PL: Znak.

Yaffe G (2010) Libet and the criminal law's voluntary act requirement. *Conscious will and responsibility: A tribute to Benjamin Libet*, 189.

Zaluski W (2009) *Evolutionary theory and legal philosophy*. Northampton, MA: Edward Elgar Publishing Inc.

APPLYING CONFIRMATION THEORY TO THE CASE AGAINST NEUROLAW

Anton Donchev

Department of Philosophy and Sociology, New Bulgarian University,
Sofia, Bulgaria

e-mail: donchev.anton@gmail.com

Abstract:

Neurolaw is the emerging research field and practice of applying neuroscientific knowledge to legal standards and proceedings. This intersection of neuroscience and law has put up some serious claims, the most significant of which is the overall transformation of the legal system as we know it. The claim has met with strong opposition from scholars of law, such as Michael Pardo and Dennis Patterson (2011), who argue that neurolaw (and neuroscience more generally) is conceptually wrong and thus perceive most of it as “nonsense” (Patterson, 2003). I expose a flaw in Pardo and Patterson’s arguments by means of confirmation theory. My main point is that Pardo and Patterson use implicit hypothetico-deductivism in their attack on neurolaw, and that we have good reasons to doubt the employment of such a model, because it faces serious theoretical problems. I then demonstrate how the alleged problems associated with neurolaw disappear if we use a quantitative probabilistic account of confirmation. I also explain why it provides a better account for the way the legal system actually works.

Key words: neurolaw, confirmation theory, hypothetico-deductivism, empirical evidence, ordinary language philosophy, legal system.

1. Introduction – the potential impact of neuroscience on the legal system

Neuroscience is a quickly developing scientific research field, starting only about a hundred years ago when scientists began dedicating time and effort to the systematic investigation of how the brain (and in particular *the human brain*) works (Gazzaniga,

2008). This rapidly expanding field brings the promise (or perhaps the threat) of making significant changes to our societal structure and to our everyday lives (Greene & Cohen, 2004). By and large, neuroscience seems to provide evidence against the philosophical notion of free will, i.e., that human beings are agents capable of making choices determined by their own volition. Thus, neuroscience casts a shadow over the notion of moral (and even legal) responsibility, which might be entailed by our possession of free will. After all, if a person’s choices are determined by events outside of their conscious control, how could we justify holding them morally (or legally) accountable for their choices?

If such persisting philosophical questions are indeed close to receiving a conclusive answer on account of present day neuroscience, this could lead to a major twofold transformation of the existing traditional legal framework, thus encompassing both criminal and civil law. On one hand, the notion of criminal responsibility would have to part ways with any element of retribution or “retributive justice” that it currently contains. No legal system would be justified in punishing an offender for their choices, if it was established that, all things being equal, they could not have chosen or acted otherwise. Thus, *preventive measures* and *treatment* would supersede *punishment* and *retribution* on the center stage of criminal law (Eagleman, 2011). On the other hand, legal decision makers would have to incorporate neuroscientific facts and findings, and learn to successfully discriminate between them. In other words, legal decision makers would have to be able to successfully distinguish between neuroscientific evidence obtained according to the scientific method and pseudo-scientific or prejudicial information.

The decision whether to allow a piece of evidence in court is usually in the hands of one person, namely the judge,¹ yet the making of such informed decisions requires the background collaboration of multiple research fields, and is in no part a trivial or easy task. In fact, the legal systems of many countries have already taken steps to implement such transformations to some extent, although it seems this process is still in its early development (Spranger, 2012). But changing the legal system on account of neuroscience deserves a word of caution: despite whatever *radical* changes the future may hold, these are today merely speculative. Therefore, it might be best to avoid falling prey to what Stephen Morse (2006) has called “brain overclaim syndrome” (BOS). In short, BOS is the tendency of people, fascinated by the discoveries of neuroscience, to make overly bold claims about its implications — claims that present day neuroscience itself is not able to support. Examples include claims about shutting down the penitentiary system and absolving every convicted felon of legal responsibility, or claims that the legal system should switch entirely from behavioral evidence to neuroscientific evidence for establishing culpability. Such claims are not direct implications of present day neuroscience and their realization is highly implausible in the foreseeable future.

2. The case against neurolaw (and its critics)

Bold demands for changes in a field as conservative as the law, whether a product of BOS or not, have inevitably met with the strong resistance of some of its practitioners and scholars. A particularly forceful line of argumentation against neurolaw, and the consistency of cognitive neuroscience as a whole, is delivered by Michael S. Pardo and Dennis Patterson. In a series of papers (e.g., Patterson, 2003; Pardo & Patterson, 2011), they have mounted an attack on cognitive neuroscience and its application to the law, which culminated in their full-length treatment of the subject in a book from 2013, entitled *Minds, Brains, and Law: The Conceptual Foundations of Law and Neuroscience*. The main argu-

ment that is used throughout this attack is based on the so-called “ordinary language philosophy” of Ludwig Wittgenstein and Gilbert Ryle. In a short and simple reconstruction, the argument goes as follows: a brain does not believe, know, feel, or lie; a person does. The activities of persons, such as rule-following, knowledge, or interpretation, cannot be meaningfully ascribed to activities of the brain. That is because our everyday mentalistic or psychological concepts like “believe,” “know,” “feel,” “perceive,” etc., are used to refer to persons, not parts of persons (such as their brains). And, according to ordinary language philosophy, “the meaning of a word is its use in the language” (Wittgenstein, 1953, §43). Hence, these words only have meanings when applied to a whole person, and applied to anything else (e.g., a brain) they are meaningless. When trying to ascribe psychological concepts to parts of persons in lieu of the persons themselves, one is said to commit the “mereological fallacy.” Pardo and Patterson also claim that many contemporary neuroscientists commit this fallacy, and that cognitive neuroscience itself is based on such misguided usage of language, thus rendering most of it meaningless.

Of course, this line of argumentation is not entirely original — it is largely based on Maxwell Bennett and Peter Hacker (2003). Bennett and Hacker, in their turn, inherit a philosophical viewpoint dating back at least to the ordinary language philosophy of Oxford in the 1960s. One of the main points of this philosophical movement is the distinction between a *personal* and a *sub-personal* level of explanation: concepts that refer and apply to persons belong to the personal explanatory level, while concepts that refer to sub-personal processes, mechanisms, or parts, such as brains and neurophysiological facts, belong to the sub-personal explanatory level. As long as this distinction is kept within the bounds of common sense, it is indeed a useful conceptual tool. However, it seems that there were attempts to radicalize the distinction between the personal and the sub-personal explanatory levels, predating Bennett and Hacker by over 40 years. Here is what Daniel Dennett warns us in 1969 about carrying the distinction too far:

In one respect the distinction between the personal and sub-personal levels of explanation is not at all new. The philosophy of mind initiated by Ryle and Wittgenstein is in large measure an analysis of the concepts we use at the

¹ In the U.S. legal system, for example, this role of the judge is entailed by their function as a gatekeeper, as provided by the Federal Rules of Evidence (Federal Evidence Review, 2015). The rules also incorporate the so-called *Daubert* standard for admissibility of expert witnesses’ testimony (cf. *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993 509 US 579 (1993)).

personal level, and the lesson to be learned from Ryle's attacks on 'para-mechanical hypotheses' and Wittgenstein's often startling insistence that explanations come to an end rather earlier than we had thought is that the personal and sub-personal levels must not be confused. The lesson has occasionally been misconstrued, however, as the lesson that the personal level of explanation is the *only* level of explanation when the subject matter is human minds and actions. In an important but narrow sense this is true, for as we see in the case of pain, to abandon the personal level is to stop talking about pain. In another important sense it is false, and it is this that is often missed. The recognition that there are two levels of explanation gives birth to the burden of relating them, and this is a task that is not outside the philosopher's province. It cannot be the case that there is *no* relation between pains and neural impulses or between beliefs and neural states, so setting the mechanical or physical questions off-limits to the philosopher will not keep the question of what these relations are from arising. (Dennett, 1969, p.95, emphasis added)

Following a similar line of thought, Sarah Robins and Carl Craver (2011) have argued against Pardo and Patterson's claim about the meaninglessness of neuroscience. Robins and Craver's main point is that even if they grant, for the sake of argument, that cognitive neuroscience abounds in misguided applications of *personal level concepts* to *sub-personal level explanations*, such is insufficient to make it nonsensical. They emphasize here the necessary relation between sub-personal mechanisms and personal level behavior. In other words, sub-personal facts, such as events and processes in a properly working brain, are necessary to bring about the behavior of the personal level (a point that is in fact granted by Pardo and Patterson). So, even if brains cannot lie, believe, or know, important information about whether a person is lying, or whether they believe or know something, might still be obtained by investigating their brain. The degree to which such an endeavor could be successful is an empirical matter and not one that can be decided "from the armchair." But the prospect of it is enough to make sound sense out of cognitive neuroscience and its application to the law. Furthermore, neuroscientific

evidence is sometimes the best or even the only evidence there is, such as in cases where any expressed behavior is essentially lacking. Such examples include cases of Millard-Gubler syndrome, also known as "ventral pontine syndrome," which may lead to a full paralysis and the inability to provide behavioral response to external stimuli, or the far more common cases where the defendant alters their expressed behavior on purpose, in order to deceive. As Robins and Craver (2011) put it: "It is not nonsense to use the best evidence at your disposal" (p.199).

It should be noted that Pardo and Patterson (2013) have moved the goalposts somewhat on account of their critics, such as Robins and Craver, and have toned down their claims for the nonsensicality and meaninglessness of cognitive neuroscience to a significant degree:

To be clear, our aim in discussing the conceptual issues is neither to challenge the scientific project nor to dismiss outright the possibility for such technology to provide probative legal evidence. Far from it. [...] We think there is much to take seriously in neurolaw, given our arguments, and our analysis does not eliminate the need to engage with difficult methodological, moral, and legal issues. The soundness of the theoretical conclusions and practical applications drawn from the experiments depend on clarity regarding the relevant concepts, and our analysis is continuous with, and intended to be a contribution to, these neurolaw projects, not an outright rejection of them. (Pardo & Patterson, 2013, p. 81)

However, their argument for the conceptual confusion in neurolaw, and neuroscience more generally, still stands. The argument is based in part on an Aristotelian conception of mind, which they define as "an array of rational and emotional powers, capacities, and abilities exhibited in thought, feeling, and action" (Pardo & Patterson, 2013, p. 44). That means "memory," for example, is the *personal level* ability to retain knowledge (p. 102). According to this view, it is meaningless to ascribe memory to the brain or to any other sub-personal part. It is also the personal level behavior that is *the measure* of what constitutes the mental property of memory. Neuroscience can at best provide *a* measurement of whether one has some mental property on a particular occasion, but is never *the* measure for the ascription of mental concepts. Furthermore, behavior

trumps neuroscientific evidence, according to Pardo and Patterson (pp. 10-12). In other words, when expressed behavior is observed, and it is inconsistent with the neuroscientific data, the neuroscience behind it has to change. Behavioral criteria are privileged in the ascription of mental concepts, because it is behavioral criteria that govern the use of mental concepts in ordinary language.

This challenge is met head-on by Neil Levy (2014). Levy argues that neurolaw is in fact not conceptually confused: rejecting the Aristotelian conception of mind, he favors a more modern conception in cognitive science, according to which the mind is “the totality of information processing mechanisms that underlie the powers and capacities on which they [Pardo and Patterson] fix” (Levy, 2014, p. 174). Thus, Levy also rejects the notion of privileged criteria for the ascription of mental concepts. This means that behavioral, as well as neural, evolutionary, cognitive psychological and even sociological evidence, may all have different weights on the ascription of mental concepts in different cases (p. 173). Just because we frequently ascribe mental concepts based on behavior (in ordinary language), we cannot readily conclude that these words merely denote behavior and nothing else. Levy holds that it is wrong to “draw metaphysics from semantics,” i.e., it is a mistake to infer “what must be the case from what it is appropriate to say” (p. 176). Indeed, it is a *non sequitur* to conclude that a brain cannot think because the word “think” is usually used to refer to an activity of a person, not to an activity of a brain.

Levy (2014) also argues that the Wittgensteinian framework of understanding is incompatible with present day cognitive science (p. 179). Wittgenstein (1953) was mainly concerned with the homunculus fallacy, i.e., that we do not explain mental properties by postulating a mechanism which has the same properties we aim to explain. Such a notion would only push the task of explaining a bit further, needlessly multiplying the entities to be explained, without actually providing any explanation of the mental properties themselves. Pardo and Patterson believe that cognitive neuroscience is committing this fallacy when, for example, it maintains that a person’s brain can store events that a person can perceive. If the brain holds such records, they say, one must see one’s own brain to access them, and also one has to remember what one’s memories are memories of (Pardo & Patterson, 2013, p. 104). This view is

based on a misunderstanding of the explanatory project of cognitive neuroscience. Neuroscience has avoided the homunculus fallacy by positing multiple simple, causal, and sub-personal level mechanisms, all of which are necessary for a person to realize any and all mental states. These mechanisms need not be self-conscious, and people need not be conscious of them, or interpret them, in order for the mechanisms to work. Both Neil Levy (2014) and Robins and Craver (2011) have made this simple point.

3. A deeper problem with the case against neurolaw

The above arguments against the attack on neurolaw, as well as cognitive neuroscience as a whole, are sound. However, there seems to be a more fundamental problem with the case against neurolaw, which concerns the method of the ordinary language philosophy used, and the claims that can be made based on it. There are two interrelated parts of the problem: a misconstrued notion of evidence and a faulty relation of confirmation between a piece of evidence and a hypothesis. Both of these problems may be considered the domain of philosophy of science – namely its subdiscipline of “confirmation theory” that is mainly concerned with the relation of confirmation or disconfirmation between empirical evidence and different hypotheses or theories.

The first part of the problem — the misconstrued notion of evidence — stems from an underlying and controvertible notion of what a philosopher’s work consists of and how it should be carried out. According to Bennett and Hacker (2003), a philosopher cannot decide on empirical matters, nor can empirical enquiry solve any philosophical problems. Philosophers are instead mainly concerned with conceptual matters of meaning and meaninglessness, sense and nonsense. They cannot decide upon questions of truth and falsity, as these are amenable to scientific enquiry. Instead, they operate on a meta-level, preceding questions of truth and falsity. On this view, when a philosopher is mistaken, she is “not even wrong”² — she is just speaking nonsense. The question is thus how philosophers investigate such matters of meaning or sense. According to Bennett and Hacker, who follow the classical answer of ordinary language philosophy, philosophers examine the use of words. However, as previously

² This famous phrase is usually attributed to the physicist Wolfgang Pauli (Peierls, 1960, p.186).

stated, philosophers cannot engage in empirical enquiry, thus rendering their examination of the use of words as based on an introspective investigation of their own “intuitions” about how these words are used, as well as their meaning.

These intuitions are then pointed out as evidence in support of hypothetical statements about the use of ordinary language words. What is more, the hypothetical statements in question are not even labelled as hypotheses or theories, instead being proffered as “rules” which govern the use of ordinary language. Anyone not following the so-called “rules” is accused of engaging in nonsense. And since cognitive neuroscience (and neurolaw in particular) dares to use words disregarding these rules, they are nonsensical as well. Pardo and Patterson subscribe to this line of thought explicitly (e.g., 2013, pp. 5-8), thus motivating to a large extent their case against neurolaw. The problem with this line of thought is summarized by Daniel Dennett: “Notice, first, that, no matter what any philosopher may say, examining the use of words is an empirical investigation, which often yields everyday garden-variety truths and falsehoods, and is subject to correction by standard observations and objections” (Dennett, 2007, p. 81).

The “use of words” in ordinary language is in fact the actual practice of real people speaking out, writing, or in any other way communicating, words (i.e., auditory, visually, or perceptible in some other way, symbols). It is not an abstract mathematical or logical process with no relation to the physical world. Therefore, the use of words is subject to empirical enquiry, and simply asserting otherwise does not make it less so. As in all such cases, when a matter is subject to empirical enquiry, it allows for the formation of hypotheses, which may be confirmed or disconfirmed by empirical evidence. However, it would not do to supplant empirical evidence with introspective intuitions, nor will it do to assert the philosopher as an expert in matters of ordinary language use, whose intuitions should count as undisputable evidence in support of her hypotheses. The nature of empirical evidence, put simply, is that it is available to more than a single person, and that it may be subjected to repeatable examination and/or experimentation. The nature of consulting one’s own intuitions is either not that different from empirical enquiry, or it can produce no evidence at all. As Dennett puts it:

One still cannot study it [grammar] without asking questions—and even if you only ask *yourself* the questions, you still have to see what you say. The conviction that this method of consulting one’s (grammatical or other) intuitions is entirely distinct from empirical enquiry has a long pedigree (going back not just to the Oxford of the 1960s, but to Socrates) but it does not survive reflection. (Dennett, 2007, p. 81)

A related problem is that, with the notable exception of experimental philosophy, no other kind of ordinary language philosophy actually tries to equip itself and use the technical apparatus needed to produce empirical evidence in support of its hypotheses. What I mean here is the use of experimental design techniques and statistical methods for analyzing data. It would not do here to object that ordinary language philosophy should not be required to meet the criteria for objectivity of natural science. If ordinary language philosophy sets as its task to fix *rules* for the use of ordinary language words as a matter, as I have argued, of empirical nature, it should use the proper methodology. This means it should first come out and clearly state its conjectures about the use of certain words in the form hypotheses. Then it should determine the population of interest (e.g., common people, or people with background in law, or in neuroscience, etc.) and the sampling unit (i.e., the sub-group of the population, which would participate in the study). Subsequently, it should gather data about how people *actually* use these words. Finally, it should employ inferential statistics to check whether its hypotheses are confirmed by the data or not. In summary, in order to settle empirical questions, ordinary language philosophy should use the best available method for such a task — namely the scientific method, and not base its conclusions on introspections about personal intuitions.

The second part of the problem is a faulty relation of confirmation between the evidence and the hypothesis in the ordinary language philosophy that is used as a basis for the attack on neurolaw. This method of confirmation would probably be neither accepted, nor recognized as their own, by most ordinary language philosophers who follow Wittgenstein and Ryle. Furthermore, it will most definitely be rejected by Bennett and Hacker, and thus also by Pardo and Patterson, who insist that their analysis in conceptual use is above empirical matters. However,

if one accepts that this analysis is in fact an empirical endeavor, and hence its hypothetical statements are subject to confirmation and disconfirmation in much the same way as every other hypothesis about the world is, one can reconstruct the main relation of confirmation, which they employ in analyzing conceptual use. It goes roughly as follows. One has a hypothesis about how some ordinary language word is used and in order to confirm or disconfirm it, one searches one's own intuitions about its use. If an intuition that matches the hypothesis is found, one concludes that one's hypothesis is definitely confirmed on the basis of such "evidence". If no matching intuition is found, and especially if there is an intuition contradicting the hypothesis, one concludes that the latter is definitely disconfirmed. This simple mechanism implicitly underlies the way ordinary language philosophy was conducted by Wittgenstein and Ryle. Although it was often masked behind phrases such as "we say so-and-so", it in fact gave no evidence about how people *actually* use a word, apart from the philosopher's *own intuitions* on the subject. It could be no other way, as this type of ordinary language philosophy was carried out "from the armchair". The same type of philosophy was inherited by Bennett and Hacker, and through them by Pardo and Patterson, and it is its simple mechanism that gives rise to overly strong claims such as defining *the rules* for the use of words in ordinary language.

The main fault of the theoretical mechanism itself is that it is hypothetico-deductive. In other words, its relation of confirmation between the evidence (or the misconstrued notion of it, in this case) and the hypothesis to be confirmed or disconfirmed, is a hypothetico-deductive one. In the simplest case of hypothetico-deductivism, we have one hypothesis, which logically entails propositions about observable facts. If we find evidence in support of these propositions, we say that the hypothesis which entails them is confirmed, or at least that it is corroborated (i.e., it has survived empirical testing). If we find evidence rejecting these propositions, we say that the hypothesis is disconfirmed or falsified. This type of confirmation is sometimes called "deduction-in-reverse" (Crupi, 2013) as truth is passed on backwards through the deductive chain - from the propositions about observable facts - to the hypothesis, which logically entails them.

Hypothetico-deductivism faces serious theoretical problems, and the discovery of these problems

has led to its demise: its basic variation has fallen out of favor in contemporary confirmation theory and has been called "naïve" (Crupi, 2013) or even pronounced to be "hopeless" (Glymour, 1980).³ One example here would be the problem of irrelevant conjunction, which consists of hypothetico-deductivism's inability to discern, for the purposes of confirmation, between a hypothesis and an irrelevant or even absurd proposition tacked to the hypothesis via logical conjunction. If some hypothesis h , in conjunction with some background knowledge k , entails available evidence e , then e hypothetico-deductively confirms $h \wedge k$ (as long as $h \wedge k$ is consistent and k does not entail e by itself) (Schurz, 2014; Crupi, 2013). However, e hypothetico-deductively confirms not only $h \wedge k$, but also $h \wedge k \wedge a$, where a is an arbitrary irrelevant proposition. For example, if my hypothesis h is "the verb 'lie' refers *only* to persons", my background knowledge k is fluency in English, and my evidence e is the observation that I always use "lie" to refer only to persons, then e hypothetico-deductively confirms $h \wedge k$. The problem of irrelevant conjunction is that e also hypothetico-deductively confirms $h \wedge k \wedge a$, where a is the irrelevant proposition "brains cannot lie". In other words, my observation about how I use a word in the language hypothetico-deductively confirms an irrelevant proposition about the capabilities of a human physiological part.

The hypothetico-deductivist reliance on formal propositional logic has the additional drawback of providing only two modes of confirmation in dichotomous relation, namely "confirmed" and "disconfirmed." It provides for no degrees of confirmation in between. And there are good reasons to believe (dating back at least to Hume's problem of induction) that no amount of inductive empirical evidence can completely and conclusively confirm a hypothesis, but it can bring a hypothesis ever closer to confirmation. This point is granted by Pardo and Patterson, who proceed to distinguish inductive evidence from "criterial evidence" (2013, pp. 8-12). The latter should be a special kind of evidence, im-

³ For its difficulties with the Duhem-Quine thesis, as well as for a good overview of all of its problems, including the problem of old evidence and Hempel's paradox, see Crupi (2013). For the paradox of conceptual difference, see Laudan and Leplin (1991, p.464); and Betz (2013). For Goodman's "new riddle of induction" and its impact on hypothetico-deductivism, see Goodman (1983, pp.67-68).

mune to the inductive weakness of never providing complete confirmatory support. Pardo and Patterson assure us that, when ascribing psychological predicates, the “criterial evidence [...] consists of various types of behavior” (p. 9). It means that behavior is the evidence that completely and conclusively confirms whether someone, e.g., perceives or believes something, because that behavior is a part of what it means to “perceive” or “believe.” This immediately raises the question: are neural states not also a part of what it means to “perceive” or “believe?” The short answer, according to Pardo and Patterson, is “no;” however, they give away little information on why this is so, except for the fact that Bennett and Hacker (2003) think that way. And, as we have already seen, Bennet and Hacker (2003) justify their belief by employing the inadequate framework of Wittgenstein and Ryle’s ordinary language philosophy, which leads to: 1) trying to answer what is essentially an empirical question — how words are commonly used — by appealing to intuitions, instead of evidence; and 2) trying to draw conclusions about how things actually are (e.g., what perception is) from how words about them are commonly used (e.g., how ordinary people use the word “perceive”).

The notion of “criterial evidence” thus suspiciously resembles a case of special pleading, and this suspicion is strengthened by the fact that contemporary confirmation theory does not recognize any such privileged class of evidence. We should note here, of course, that contemporary confirmation theory deals with relating empirical facts to hypothetical statements, and Pardo and Patterson admit of making no such relation, insisting instead that they deal with conceptual issues above empirical enquiry. If we recognize, however, that “criterial evidence” is either just another type of empirical evidence, or no evidence at all, we would see why it might be problematic to ground very strong beliefs on it.

Unfortunately, the lack of degrees of confirmation in hypothetico-deductivism leaves the (in effect false) impression that, once confirmed, a hypothesis is one hundred percent confirmed. This impression seems to be at the core of the strong claims about defining *rules* for the use of ordinary language words. After all, if you find your hypothesis a hundred percent confirmed, without any room for possible doubt, why not promote it to the rank of a rule? Furthermore, if one does not follow the said rule, one cannot be anything but mistaken, as it is impossible the rule itself to be false or inaccurate. And if

the rule is about the use of a word, which fixes its meaning, yet someone does not obey it, his/her use of this word must be meaningless — there is no other way it could be.

4. The problems of neurolaw and a quantitative probabilistic account of confirmation

Let us now consider a quantitative probabilistic account of confirmation, in place of the hypothetico-deductive one. Probability theory confers the advantage of providing quantitative models with multiple (in theory infinite) degrees of confirmation in between the extreme cases of “confirmed” and “disconfirmed.” This allows for the fine tuning of the “confirmed” or “disconfirmed” status of a hypothesis, as well as helping to avoid overly strong claims such as promoting a hypothesis to a rule. Provided one does not begin with radical beliefs about their hypothesis, which leave no room for doubt, they always stand the chance of learning something new about their hypothesis, on the basis of new empirical evidence. In other words, a hypothesis will never reach one hundred percent confirmation, but may come asymptotically closer to it as more and more evidence in its support is found. It seems that such a view would make one hesitant to promote one’s hypothesis to the rank of a rule, especially based on a few personal intuitions.

The second thing to note is the possible role of the intuitions themselves in a quantitative probabilistic account of confirmation. Even if we accept, for the sake of argument, that some intuitions may constitute a legitimate form of evidence, they may still not be allowed to have weight on the confirmation or disconfirmation of the hypothesis in our case. That is because the same intuitions have already played a role in the formation of the prior beliefs about the hypothesis. In most cases, the hypothesis about how some ordinary language word is used is in fact a pet theory of the ordinary language philosopher, such as, for example, “psychological concepts (e.g., “believe”, “know”, “feel”, “perceive”) refer to persons and not to sub-personal processes, mechanisms or parts”. Such a hypothesis is already informed by, and even based on, the philosopher’s own intuitions. It will not do to use the same intuitions again to support it, otherwise that would be circular reasoning. Therefore, either the hypothesis should be intuition-free, which could be hard in the case of ordinary language philosophy conducted “from the arm-chair”, or the intuitions which form the prior beliefs

may not count again as supporting evidence for its confirmation. It will thus become necessary for one to seek supporting evidence elsewhere, and, as I have proposed, the best way to find it is to investigate the actual use of words by real people.

When evaluating the degree of confirmation of a hypothesis about the use of psychological concepts, one should either use all the evidence at one's disposal, or provide good reasons why some of it may be excluded from the analysis. Otherwise, the result may become skewed for no good reason, thereby failing to provide the measure of objectivity that is expected. If one uses all available evidence, it seems one would have to account for the use of psychological concepts in *neuroscience* and even in *neurolaw*. Pardo and Patterson's hypothesis that psychological concepts refer only to persons, and may not refer to parts of persons, such as their brains, would be disconfirmed (i.e., rendered largely improbable on the basis of such evidence). And even if one finds good arguments to exclude evidence of the use of psychological concepts from cognitive neuroscience, it is still questionable whether Pardo and Patterson's hypothesis would hold. It may well be found that outside cognitive neuroscience (and perhaps even in ordinary language), these words do not refer to persons alone but have various non- or sub-personal referents:

It is not just neuroscientists; it is computer scientists (and not just in AI), cognitive ethologists, cell biologists, evolutionary theorists ... all blithely falling in with the game, teaching their students to think and talk this way, a linguistic pandemic. If you asked the average electrical engineer to explain how half the electronic gadgets in your house worked, you'd get an answer bristling with intentional terms that commit the mereological fallacy—if it is a fallacy. (Dennett, 2007, p. 87)

In summary, if we recognize that Pardo and Patterson's "criterial evidence" is either a type of empirical evidence, or no evidence at all; and, if we explicitly use a quantitative probabilistic account of confirmation in place of the faulty implicit hypothetico-deductive one, we would get no "rules" for the use of ordinary language words, hence no conceptual confusion and "no nonsense neurolaw" (Robins & Craver, 2011). Then no one could be rightfully accused of inadequate use of language (unless shown otherwise on the basis of empirical evidence, i.e., on

properly gathered and analyzed data on the use of language).

5. Conclusion: "evidence" and "confirmation" in the US legal system

The concept of empirical evidence that I have argued for thus far, as well as the probabilistic account of confirmation proposed above, seem to be in agreement with the concept of evidence and the mechanism of confirmation in the US legal system. In any case, they bear more similarities to it than the notion of evidence based on Wittgenstein and Ryle's ordinary language philosophy, and its implicit model of confirmation. A quick review of the Federal Rules of Evidence (Federal Evidence Review, 2015) would convince us that the word "evidence," as used in a US court of law, is incompatible with the meaning of personal intuitions about the use of words. For example, a US court of law would not conclude that a certain corporation X could *not* have unlawfully appropriated funds from its clients, just because "unlawful appropriation" and its more common synonym "stealing" are words that traditionally apply to persons, not abstract entities, such as corporations. Thus, terms like "corporate fraud" would be nonsense and claims that "corporation X stole your money" would be met with derision by the law — corporation X cannot have stolen your money, as corporations do not "steal" — persons do. However, with the rise of modern corporations the lawmakers did not forge the laws about corporate fraud based on their intuitions about how ordinary language uses the word "fraud". They simply extended the meaning of the word "fraud" to abstract entities of the new type "corporation" and this extension slowly seeped into ordinary language as well. What is more important here, concerns about the ordinary language use of certain words would not be permitted as evidence in court — that is why conclusions of the type that the corporation cannot have stolen your money, as corporations cannot "steal", sound absurd. Although the law's understanding of "evidence" is distinct from the one of science, they have at least that much in common — both are empirical and subject to empirical enquiry.

As for the law's account of confirmation, it seems that it is a quantitative probabilistic one. For example, here is how evidential relevance is defined by the Federal Rules of Evidence:

Rule 401. Test for Relevant Evidence

Evidence is relevant if:

- (a) it has any tendency to make a fact more or less probable than it would be without the evidence; and
- (b) the fact is of consequence in determining the action.

(Federal Evidence Review, 2015, Article IV)

The estimation of a piece of evidence's relevance to a case is based on a probabilistic mechanism of confirmation, which is already quite familiar in confirmation theory. If an evidence makes a fact more probable than it otherwise is, the evidence counts in support of the fact, thereby increasing its degree of confirmation and making it relevant for the case. In a similar way, if an evidence makes a fact less probable than it otherwise is, we would say that the evidence disconfirms the fact to some degree, and that again makes it relevant for the case. And if the evidence does not affect the probability of the fact, it does not change its degree of confirmation and is therefore irrelevant to the case, unless it is of consequence in determining the action of interest to the court.

If the legal system employed a notion of evidence or a confirmatory mechanism similar to those implicitly used by the ordinary language philosophy of Pardo and Patterson (2013) and Bennett and Hacker (2003), this would potentially lead to an absurd scenario — imagine a legal system, which considers appropriate for its judges to decide upon real life cases, based on their own intuitions on how some words are used in ordinary language. The transformations that may await such a legal system on account of cognitive neuroscience would pale in comparison to its more pressing problems. In fact, even the possible obsolescence of the notion of retribution (which is often considered to be one of the more radical claims of neurolaw) seems like a trivial prospect compared to the changes that would be needed in such a legal system in order to restore to it justice, or even common sense.

References

- Bennett M, Hacker P (2003) *The philosophical foundations of neuroscience*. Oxford, UK: Blackwell Publishing.
- Betz G (2013) Revamping hypothetico-deductivism: A dialectic account of confirmation. *Erkenntnis*, 78(5): 991–1009.
- Clarke C (2014) Neuroeconomics and confirmation theory. *Philosophy of Science*, 81(2): 195–215.
- Crupi V (2016) Confirmation. In *The Stanford encyclopedia of philosophy*, ed. E. N. Zalta, (Winter 2016 Edition). Retrieved from <http://plato.stanford.edu/archives/win2016/entries/confirmation/>.
- Dennett D (1996) *Content and consciousness*. London, UK: Routledge & Kegan Paul.
- Dennett D (2007) Philosophy as naive anthropology: Comment on Bennett and Hacker." In *Neuroscience and Philosophy: Brain, Mind, and Language*, eds M Bennett, D Dennett, P Hacker, J Searle. New York, NY: Columbia University Press.
- Eagleman D (2011, July/August) The brain on trial. *The Atlantic*. Retrieved from <http://www.theatlantic.com/magazine/archive/2011/07/the-brain-on-trial/308520/>.
- Federal Evidence Review, "Federal Rules of Evidence," (2015) December 20, 2017. <http://federalevidence.com/rules-of-evidence>.
- Gazzaniga M (2008) The law and neuroscience. *Neuron*, 60(6): 412–415.
- Glymour C (1980) Hypothetico-deductivism is hopeless. *Philosophy of Science*, 47(2): 322–325.
- Goodman N (1983) *Fact, fiction and forecast*, 4th ed. Cambridge, MA: Harvard University Press.
- Greene J, Cohen J (2004) For the law, neuroscience changes nothing and everything. *Philosophical Transactions: Biological Sciences*, 359(1451): 1775–1785.
- Laudan L, Leplin J (1991) Empirical equivalence and underdetermination. *The Journal of Philosophy*, 88(9): 449–472.
- Levy N (2014) Is neurolaw conceptually confused? *The Journal of Ethics*, 18(2): 171–185.
- Morse S (2006) Brain overclaim syndrome and criminal responsibility: A diagnostic note. *Ohio State Journal of Criminal Law*, 3(2): 397–412.
- Pardo M, Patterson D (2011) Minds, brains, and norms. *Neuroethics*, 4(3): 179–190.
- Pardo M, Patterson D (2013) *Minds, brains and law: The conceptual foundations of law and neuroscience*. Oxford, UK: Oxford University Press.
- Patterson D (2003) Fashionable nonsense. *Texas Law Review*, 81(3): 841–894.
- Peierls R E (1960) Wolfgang Ernst Pauli. 1900–1958. *Biographical Memoirs of Fellows of the Royal Society*, 5: 174–192.
- Robins S, Craver C (2011) No nonsense neurolaw. *Neuroethics*, 4(3): 195–203.

Schurz, G. (2014). *Philosophy of Science: A Unified Approach*. London, UK: Routledge.

Spranger T (ed.) (2012) *International neurolaw: A comparative analysis*. Berlin, Germany: Springer.

Wittgenstein L (2009) *Philosophical investigations*, revised 4th edition (G E M Anscombe, P M S Hacker, J Schulte, Trans.). Hoboken, NJ: Wiley-Blackwell. (Original work published 1953).

FREEDOM, RESPONSIBILITY, AND JURISPRUDENCE

Hari Narayanan V

Department of Humanities and Social Sciences, IIT Jodhpur, India
e-mail: hari@iitj.ac.in

Abstract:

This paper seeks to argue that advances in the study of freewill and responsibility are directly relevant to jurisprudence. Following Daniel Dennett attempts to discredit the existence of freewill with the help of experiments can be checked by arguing that freedom should be understood as something that has evolved over time rather than being a pre-existent feature of our species. The major function served by freedom is to ensure responsibility for actions. This understanding of freedom as something that evolved to enhance responsibility suggests that freedom can be developed further. This can be understood as enhancing the ability to follow social norms by overcoming factors that limit responsibility. Jurisprudence has to take into account the ability to follow norms as a variable, even within the category of adults, and treat violations accordingly. Further, efforts to enhance the capacity to be free from habitual reactions need to be made part of education, and the state has to focus on this aspect without which the task of ensuring adherence to law of citizens will remain incomplete.

Key words: freedom, free will, responsibility, jurisprudence.

Introduction

The nature of free will and its impact on the notion of responsibility have been a matter of debate for long. The debate is often centered around voluntary actions in order to find out whether those actions are chosen freely, so as to make the agent truly responsible for the action. If everything is caused by something else, then it follows that even the choice that we make is the effect of something and, therefore, cannot be treated as free. This is often referred to as determinism, a thesis according

to which the choices available for action are actually determined by factors outside of us. That is to say, even if we tend to think that one particular course of action is our free choice, it is not actually the case because, given the history of the universe and the laws of nature, it would not have been possible for us to choose any other option. At the same time, it appears that we are free in exercising a choice as we do not feel any constraint in normal circumstances, nor are we able to identify any cause behind the choice. This debate is known as free will versus determinism. There have been attempts to find a middle path by arguing that there can be a notion of free will compatible with determinism.

Following Dennett, I shall argue in this paper that freedom *qua* an evolved feature of human beings helps to ensure responsibility for particular actions. Additionally, any evolved feature is subject to further progress; therefore, freedom can be further enhanced, such that legal systems must take into account the different levels of freedom within adult populations. Finally, since freedom too can be seen as something to be developed further, like knowledge, education can play a significant role in the enhancement of freedom and responsibility.

In the beginning, I shall discuss Benjamin Libet's studies and some attempts to salvage freewill from the implications of these studies. In subsequent sections, I shall discuss how freedom can be understood as something evolved to ensure responsibility for actions, and how enhancement of freedom can help in the endeavour to make actions more in tune with legal and moral norms. I shall conclude by arguing that legal systems can and need to take into account the capacity of freedom as a variable effecting changes in the way responsibility is ascribed, with remedial measures for violations being suggested.

Libet's Studies and Free Will

Some recent advances in the field of cognitive science directly impinge on the debate between freewill and determinism. Of these, Benjamin Libet's experiments are the most well-known. Libet's experiment consisted in asking a subject to move their finger and recording the readiness potential of the act in the brain (Libet, 2004). In addition, the subject was asked to report exactly the moment at which he becomes consciously aware of the intention to act. What the experiment showed was that the subject becomes consciously aware of the decision to act 150-200 milliseconds before the act, but the brain process to do the act started 550 msec before the act. What this implies is that the actual physical mechanism underlying an action precedes the occurrence of conscious thought related to the act. At the same time, human beings who are not aware of what actually goes on in the brain end up thinking that it is the conscious thought that has caused the act. This underlies the common belief in free will.

If the actual causal process precedes the occurrence of conscious thought, then the function of conscious thought, as the psychologist Daniel Wegner observes, can only be that of alerting the organism that a particular act is going to be performed (Wegner, 2002). In cases of involuntary actions, such information about the onset of action is not available to consciousness. But this is not the way the difference between voluntary and involuntary acts is understood because voluntary actions are supposed to be acts which human beings perform on the basis of their freedom of will. This free will even assumes the status of something supra-causal because it is considered to be uncaused by any natural process.

Libet's own stand in the light of his findings has not been in favour of debunking free will (Libet, 2011). He argues that even after unconscious activities initiated an action there is still the possibility of conscious veto. That means conscious thought can decide not to perform the action. This is because, in his experiments, conscious will appears about 150msec before the movement occurs. This time is sufficient to veto the movement resulting in its non-execution. In fact, he did conduct experiments where subjects could veto an act, planned for performance at a prearranged time, between 100 to 200 msec before the time set. It was also recorded that readiness potential preceded the veto, thereby showing that the subjects were

preparing to perform the action. Libet argues that there is no evidence (as of now) to connect any unconscious brain activity to the veto decision. This is to say that there is no unconscious brain activity detected that precedes the decision to veto the event. Regarding the debate between free will versus determinism, Libet holds that, whereas there is no hard evidence for both positions, the intuitive picture of humanity is in favour of free will and, therefore, to be preferred.

It is pertinent to note that Libet's argument is limited to the claim that there is no logical imperative in any mind-brain identity theory that requires specific neural activity to precede and determine the nature of a conscious control function. He adds by saying that there is no experimental evidence against the possibility that the control process may appear without prior unconscious processes. This is far away from demonstrating that conscious thought, on its own, is causally effective. It is clear that Libet has to appeal to the intuitive picture to salvage free will, which does not help much given the fact that, as science progresses, it often happens that manifest images are discarded in favour of scientific images. Further, the self-control required to veto is difficult to master. There are umpteen occasions where the attempt to control or veto results in the proliferation of actions intended to control. This indicates that it need not be the conscious thought *per se* that plays the causal role either in performance or in vetoing, but something else.

Salvaging Free Will

There have been attempts to protect the notion of free will from the experimental findings that Libet made. For instance, Alfred R. Mele has given elaborate justifications to hold on to free will (Mele, 2009). He argues that there is powerful empirical support for the claim that, at times, the conscious decision to perform an action has a place in a causal explanation of that intentional action. Mele cites one study done on distal intentions (Mele, 2009, p.135). In this study, subjects reported strong goal intentions to perform breast self-examination during the next month. Among the subjects, one group had been induced to form strong implementation intentions by specifying the date and time of the action whereas the control group was not induced to do so. Whereas everybody in the former group performed the action, only fifty three percent did so in the case of control group. According to Mele, this shows the effect of

conscious intention in the actual performance of the action. But what this really shows is that conscious intention can have an impact on factors which may remain unconscious, but which are causally responsible for an action. This means that the possibility remains open for actual causal roles being played by unconscious factors where consciousness serves primarily a flagging function.

Mele tries to counter the objection within the cases he considers, stating that both unconscious and conscious thoughts can lead to action. He gives an example where both mother and father can take the child to school on one day, and then the mother took the child on another day; but, this does not mean that mother's action is not the cause of child reaching the school. The example is clearly misleading. There is already an assumption that both conscious and unconscious factors can cause an action. Unless it is proved that conscious thought exists without being caused by unconscious factors, the analogy clearly breaks down. There is a clear parity between the two in the example to which he refers, but such a parity cannot be assumed without sufficient evidence in the case of conscious and unconscious thoughts. Mele attempts to salvage his position by referring to the *sui generis* nature of thought but this fails primarily because we are far away from having a clear understanding of what thoughts are. Further, the principle of parsimony demands us not to multiply entities without necessity and therefore it is better not to ascribe any *sui generis* status to conscious thoughts.

In a nutshell, Mele has to marshal more arguments in support of his position. His attempt is not in the right direction as he tries to defend the *sui generis* nature of thought for which there is hardly any prospect of empirical support. Further, the line of reasoning that Mele pursues is easily subject to the counter arguments eloquently given by Galen Strawson (Strawson, 1994). Strawson's argument is centered on the premise that nothing can be *causa sui* (i.e. the cause of itself), and to be truly morally responsible for one's action one has to be *causa sui*. Therefore, it follows that nothing can be truly morally responsible. Whereas the argument of Mele has to meet this kind of objections it is pertinent to note that the line of thinking that Strawson pursues is based on the assumption that the self is something with rigid boundaries, hence why nobody is truly responsible: many of the person's mental traits are not the result of his or her choice. But the assumption that the self is something with rigid

boundaries is clearly a case of pre-reflective understanding and, therefore, subject to further inquiries.

The prototype approach to categories takes cudgels against the notion of a category having rigid boundaries by pointing out the fuzziness inherent to categories: psychological studies show that humans tend to understand categories by means of a prototype (Rosch, 1975). For example, if we want to determine whether an item, such as a carpet, does or does not belong to the category of furniture, we compare the carpet with some prototype of "furniture" that we have in mind. Such a process will result in cases that do not clearly belong to a category, and this is what the prototype approach to categories advocates — that there need not be strict criteria for category membership. This line of thought suggests that the self has fluid boundaries, a view which opens up avenues to look at responsibility as a graded notion. That is to say, if the boundaries of the self cannot be demarcated with precision, then there may not be any such thing as complete responsibility. This makes Strawson's argument tangential to responsibility understood in both legal and moral senses.

Freedom and Evolution

Since both Libet and Mele fail to salvage free will, what is left is to argue to the effect that the notion of free will is not applicable to the phenomena that Libet's studies deal with. This is the route that is taken by Daniel Dennett and Shaun Gallagher. Dennett, in his book *Freedom Evolves*, argues that unconscious processes preceding conscious decision are part of the system that he calls "the person." He writes: "once you distribute the work done you have to distribute moral agency as well ... you are not out of the loop ... you are the loop." (Dennett, 2003, p.242)

Dennett's argument proceeds by pointing out that free will is something that evolves like most other features that exist on the planet. Just like the planet developed to support life, and different life forms subsequently evolved, freedom too evolved as a feature of our species, in what Dennett calls the conceptual atmosphere of intentional action, planning, punishment etc. (Dennett, 2003, p.10) Freedom may appear from our perspective to be a phenomenon that is not the result of evolution, or as something that is ahistoric; but, this is in fact not the case. The appearance of freedom helps in the shaping and enhancing of human life. It is real,

Dennett argues, just like music and money are real. Such an account of freedom, he contends, can establish the role of moral education, and even the role of religion as a method to sustain morality.

Dennett combines determinism with freedom by pointing out that determinism does not imply inevitability, nor does indeterminism necessarily entail freedom. According to him, concepts of necessity, possibility, and causation are neutral with regard to determinism or indeterminism. A deterministic world can have many options. It is well known that systems can be chaotic and unpredictable while remaining within the realm of a determined universe. But the objection that free will enthusiasts have to this form of argument is this: if chance is taken as a factor, then there is no scope for responsibility. Dennett points out that such criticisms stem from a misunderstanding of what free will amounts to within the human realm. To understand freedom and, consequently, responsibility as evolved features is to accept that most life forms cannot be ascribed this feature. Only when we try to understand action in terms of reasons or intentions does freedom have any relevance. It is in this sense that Dennett says freedom is something that is evolved rather than being an essential feature of universe. Moreover, this freedom does not require any independence from physical causes, instead only requiring the ability to act in accordance with reasons. As long as the ability is intact and developed in a human being, he or she can be said to be held responsible for their actions. It is needless to say that the evolution of moral thinking presupposes the ability to give reasons. As the human mode of life became more complex with the onset of agriculture, such changes must have happened and freedom as a concept (or meme) became acceptable in due course of time.

Gallagher, in a similar spirit to Dennett, argues that framing the question of freewill in terms of sub-personal processes is misguided because freewill cannot be squeezed into time frames of 150-350 msec, and it is applicable primarily to intentional actions, not to simple motor processes (Gallagher, 2005). According to him, the larger time frame to which the notion of free will is applicable includes previous moments and conscious recognition of those things in terms of which an action is done. It is clear that conscious recognition of things includes different kinds of knowledge, inclination etc. That is to say, only when what is commonly called the character of a person is involved in an action, the

question whether the action is done out of free will is applicable. This means that the kind of actions studied by Libet in his experiments are not even candidates for free will. Activities of neurons such as firings and movements of muscles are all at the sub-personal level, whereas free will is ascribable only to actions that are at the personal level (understood in terms of intentions or purposes). Gallagher points out that, similar to how neurons that generate the color red are not themselves red, motor movements carrying freely chosen intentional actions themselves are not to be called freely chosen.

The argument that Gallagher makes proceeds by pointing out that conscious deliberation involves memory and knowledge, and it has real effects on behavior. It can influence, guide actions, and can function as the control mechanism. This is not to say consciousness can always exert a control. Further, this does not mean that conscious thought is *sui generis* or it is not the result of any unconscious process. It can be the case that many aspects of the production of an action take place due to unconscious activity; but, as long as consciousness has a role to play in it, an action can be considered intentional. It is the whole system, including physical and social environments, that exercises free will and is therefore responsible for action. Conscious thinking plays a role in this as it is the result of past knowledge and can direct further courses of action. At the same time, conscious thinking is not the sole factor in intentional actions; thus, there is no need to posit any supra-causal power to thought. Free decisions, according to Gallagher, do not occur by chance or randomly. Rather, they are part of the causal order.

The point that Gallagher makes can be elucidated and developed further by looking into the way the word freedom works in language. Freedom is normally as being from something. This means that someone being free to act or choose entails that they possess freedom from something. It only means they are not forced by anyone's command or social customs or habit to act in a particular manner. At the same time, it hardly means that they are free from natural causes, whether it be at the neurological or the atomic level. Now, the more interesting question is whether it is appropriate to punish someone if they do something as a result of natural causes, which is to say that nobody blames anyone for a natural calamity such as thunderstorm. So, why should we cast blame on someone if they perform an action because neuronal processes influenced them to do so?

There are two ways to address this concern. First, the system of reward and punishment is part of the institutions which have evolved, all of which presuppose personhood and responsibility. Neurons are at the sub personal level but responsibility lies with the whole person. Further, punishment goes with responsibility but reward need not be, as in a case like the lottery. That is to say, the system implicitly recognizes that chance can be an element in the genesis of human action, and there is no true responsibility in the Strawsonian sense of the term. Secondly, since both freedom and responsibility operate at the level of the whole person, if we follow the accounts of Dennett and Gallagher, then lack of choice at the sub-personal level does not diminish freedom and responsibility at the personal level. It is pertinent to note that formation of personality is a process that involves many environmental factors, including the sources of ideas and attitudes which go into its constitution. Nowhere can we certainly say that a particular person is solely responsible for their character traits. At the same time, a person is expected to make use of the ability to act in accordance with reasons; and, as long as they fail to make use of it, they can be held responsible for the failure.

Thus, when a person is held responsible for an action and punished for a crime, only the following factors need to be taken into account: whether he or she could have acted differently had their thought process, not at the particular moment but as part of their 'personality', been different. This can be said to reduce morality to what an average person can do, but that does not diminish the role of morality. Polluting air or water is not considered to be immoral in our times because most people in the industrialized world either do it or benefit from it. But it is possible that after some decades when pollution becomes extremely serious, polluting will be treated as immoral. In other words, adherence to moral norms is not beyond the ability of an average person. Were this not the case, the norms would not have evolved at all. Such is the case with most of the laws legislated for citizens to follow.

Freedom for Responsibility

When an account of free will is given in terms of the whole person rather than in terms of a particular process, a whole avenue opens up to understand the notion of responsibility. It is the person who is responsible and not any sub-personal process within the person. In this respect, freedom is

like meaning: the particular cognitive process at the sub-personal level, such as visual processing or language comprehension, can be said to take place without understanding meaning. But, the whole person is said to understand meaning. In other words, meaning comprehension is a personal level attribute, and various sub-personal processes which eventually result in meaning cannot be said to understand meaning. Similarly, freedom is a property of the whole person though the various sub-personal processes underlying the actions of the person need not be free. This gives room for the kind of phenomena studied by Libet to exist without undermining the notion of freedom as such.

Since the practice of giving reasons presupposes responsibility — viz. the question “Why did you do this?” cannot be answered unless you think that you have done something and are thereby responsible for it — it can be argued that freedom has evolved for the purpose of ensuring responsibility. As per the evolutionary account of freedom given by Dennett, it is the communicative function that is primarily responsible for the emergence of self-consciousness which, in turn, leads to the demarcation between my actions and other's actions, or between actions and events. That means, when the ability to communicate flourishes with the appearance of human language, the capacity to monitor one's actions is developed because the lack of monitoring what one is doing makes it difficult to report what is going on. With a clear demarcation between one's and other's actions, the notions of freedom and responsibility come into being. Hence, it can be contended that freedom has evolved with the purpose of ensuring responsibility. In absence of the belief that one is free in doing an action, it is not possible to feel responsible for it. Just like money exists based on the shared belief of a community, so too freedom exists based on certain beliefs and related practices. Putting it another way, freedom is not to be understood as illusory or inconsequential, instead being real and playing a significant role in everyday life.

It has been pointed out that the concept of will, as it is understood in our times as a power to freely decide, has not been present in human thinking for long. For instance, Aristotle's notion of voluntary action does not make any appeal to will (Cary, 2007). Voluntary actions are explained in terms of a syllogism involving reference to belief and desire. Action is said to directly follow from belief and desire without requiring an intermediary called will.

For instance, if I have the desire to eat, and if I believe that the particular thing found before me is edible, then I end up eating it. The point is that there is no additional thing called “will” that is posited between the judgment and the action. In the case of involuntary actions, such an explanation is not feasible as there are no associated beliefs or desires that the person holds. The concept of will in its modern sense is traced to the work of Augustine of Hippo. According to Richard Sorabji, the concepts of freedom and responsibility were separated from the concept of will power, but it is in Augustine they began to cluster. This is where the origin of the modern concept of will lies (Sorabji, 2004). In the modern sense of the word, unless one is free and therefore responsible, one cannot be said to have exercised will.

When a reference is made to the origin or evolution of the concepts of freedom and responsibility, it is important to note that such phenomena exist only with reference to the concept concerned. This is to say that there is no point in saying that the phenomenon of freedom exists independent of the concept of freedom. This was made clear by Dennett drawing an analogy with the concept of morality: It cannot be said that morality exists independent of the concept of morality, and this is so with things like freedom, will, etc. (Dennett, 1986). While there is no point in talking about morality before the very origin of the concept of morality, once the concept becomes deeply entrenched it assumes the form of a character trait and appears to be independent of the concept. A similar argument can be made for freedom. Further, just like people differ greatly as far as their moral traits are concerned, freedom too appears according to degrees. Many of the traits can be strengthened with practice and related techniques, such that morality is taken to be something that can be inculcated and strengthened. So is the case with freedom. This understanding of freedom is directly relevant for jurisprudence.

Law and Freedom

Two case studies that are often taught in law schools to shed light on what it is to be a voluntary action will help in explaining how freedom can and needs to be strengthened (Yaffe, 2011). One case is that of Huey Newton, a man shot in the abdomen who moments later fired a gun at and hit the police officer who first shot him. He testified that he had no memory of firing his gun and this was supported by

medical testimony to the effect that people suffering from severe physical trauma could undertake such movements and have no memory. He was acquitted on grounds that he was not conscious, thus rendering the act as not voluntary. In another case, Mrs. Cogdon saw a dream in which her daughter was attacked by an enemy. During the dream she tried to attack this enemy with the help of an axe in order to protect her daughter, but she actually ended up killing her daughter asleep beside her. She was acquitted on the grounds that the bodily movements which resulted in the death of her daughter were not voluntary because the mental activity guiding those movements was not conscious.

Both of these cases illustrate how responsibility is not ascribed due to a lack of consciousness underpinning the action at the time it was committed. At the same, even in criminal acts that are not premeditated, responsibility is determined because the action is still influenced by knowledge and beliefs the person holds. That is why a different person would have reacted to the same situation in a different manner; or, there is a difference between wagging one's finger and murdering a human being, as far as free will and responsibility are concerned, even if both involve muscle movements and did not involve much conscious deliberation. Even if the murder took place “spur-of-the-moment,” it shows the lack of respect that the person has for human life implicit in their beliefs. These aspects are included in what is commonly called the character of the person.

The two case studies discussed above make it clear that punishment presupposes the ability to be conscious, thereby implying that the person could have acted otherwise. That is to say, a person is held responsible for an action if and only if he or she, while performing the action, was in a position not to do it. It may be objected that this excludes cases when a person is intoxicated or under the influence of drugs, etc. In such cases it can be easily defended that the person could have avoided getting intoxicated at the first instance; but, when a person becomes mentally unstable, laws in general do not prescribe any penalty, though they would prescribe that the person undergo treatment. Similar are cases of juvenile delinquency when a person is not treated to be grown enough to be held fully responsible.

As with most categories, fuzziness crops up in many cases despite there being clear cases when a person can and cannot be held responsible. So, when a person acts impulsively and slaps someone, or

when a person is carried away by herd mentality and carries-out cold blooded acts which he or she would not have done had it not been for the mob frenzy, the extent to which the person is responsible becomes difficult to establish. Apart from such usual cases there is the more profound problem as to how far a person can be held responsible when actual causes of an action are beyond the control of conscious thought.

The account of freedom and responsibility as something evolved in the context of related psychological phenomena paves the way for the dissolution of such problems. At the same time it calls for a new look at the way punishment is meted out. Freedom is a feature that evolved to ensure that a person is responsible for his or her actions. This in turn implies that freedom can be developed further so as to make a person more responsible for their actions. Further, it clearly shows that, just like other traits, freedom varies from person to person and across particular times and contexts.

If awareness of an action and its consequences at the time of performance is the key to measure responsibility, as shown in the legal examples discussed earlier, then it follows that any effort to improve awareness *per se* has to be adopted wholeheartedly: if freedom is evolved to ensure responsibility, then freedom can be enhanced by limiting those factors which preclude responsibility. The most important factor that precludes responsibility is lack of awareness. Even if we follow Aristotle's notion of voluntary action where action follows from desire and belief, it is important to work at the level of belief and desire to enhance the moral quality of actions. Here too the importance of enhancing awareness cannot be underestimated.

With the centrality of awareness for responsibility, increasing awareness will lead to enhancement of responsibility. It is established in a series of psychological studies that much of cognitive processing takes place in the unconscious, and a person may not have much awareness of the underlying motivations behind his or her action. Further, as Eagleman points out, poor impulse control is a characteristic of the majority of law breakers (Eagleman, 2011). So, even voluntary actions are not fully voluntary in the sense that the person performing the action has full conscious control over it. It has been suggested that non-judgmental awareness can serve to reduce impulsivity, making the behavior more self-endorsed (Brown, 2007). It has to be non-judgmental because

the tendency to judge instantaneously underlies impulsive acts. This suggests that cultivating the ability to control impulses amounts to enhancing freedom and responsibility, both of which can improve the rate of adherence to rules and regulations.

Education in general — and civic education in particular — needs to pay attention to the task of enhancing freedom by those means which strengthen awareness, and the legal system has to take note of the fact that responsibility is a graded notion. It is important for the legal system to realize that helping a person to achieve greater responsibility for his or her behavior is a collective requirement. Our education system can do much in this regard: a recent study of rapists in a jail in India by a doctoral student came to the conclusion that most of the rapists are not even capable of understanding the meaning of the word “consent” (Pandey, 2016). A person who does not entertain the concept of consent cannot be said to be fully aware of the nature of the act of violation and its consequences. Accordingly, such a person cannot be said to have the intention to violate in the same degree as that of someone who possesses the concept of consent.

Even though jurisprudence certainly takes laws into account as still evolving, the ability of humans to abide by them is not normally understood as something that evolved. Though laws distinguish between children and adults, and they admit certain individuals (such as mentally disabled persons) as not having the ability to realize consequences, the question vis-à-vis how far an adult is able to regulate his or her behavior is not seriously taken into account while formulating laws. It is pertinent to note that the special courts dealing with juveniles and aiming at rehabilitating them have originated only in the last couple of centuries. The English Common Law doctrine “Malice supplies the age” used to underlie many of the judicial pronouncements: per the doctrine, if a child can be treated to have malice, then he or she can be treated as an adult in the context of a crime (Blackstone, 2016). Thus, for example, a child who hid himself after committing a heinous crime is said to be able to distinguish between good and bad, and is therefore liable to be prosecuted.

The concept of *mens rea* of criminal law assumed its present form over centuries of modification. According to Chesney, recognition that intention had to be present for responsibility started appearing only in twelfth century, which is to

say that criminal culpability was ascribed — even if there was no intention to commit the act — prior to the twelfth century (Chesney, 1939). Chesney calls this transition from the primitive liability concept to one of general blameworthiness. This was a gradual process, and it is commonly accepted that intention forms a major factor in determining responsibility in our own time. But this has not developed into consideration of the ability to manage and revise one's intentions among adults. It is time the concept of varied responsibility among adults gained currency within legal discourse. The ability to self-regulate varies widely among adults due to multiple factors, and taking this variability into account in judicial thinking is a natural extension of both the doctrine 'malice supplies age' and the concept of *mens rea*. This is indeed a task that arises out of contemporary understanding of human nature which is the result of multi-disciplinary studies on cognition. Freedom *qua* an evolved capacity to ensure responsibility provides the much needed theoretical framework for undertaking efforts to bring legal discourse more in tune with the growing body of knowledge regarding human nature.

Conclusion

Human freedom is something which has evolved over time, with the ability to give and act in accordance with reasons, and this leads us to understand the function of freedom in terms of the enhancement of responsibility. Thus, freedom is supposed to enhance the voluntary nature of voluntary actions and make the person more responsible for his or her action and towards society. Civic education requires further enhancement of freedom because attempts to follow laws and rules can go astray in its absence. Jurisprudence must similarly take into account the ability of a person to follow norms, and that efforts can be made to enhance the same.

References

- Blackstone W, Paley R (2016) *Commentaries on the laws of England: Book the fourth*. Oxford: Oxford University Press.
- Brown K W, Ryan R M, Cresswell J D (2007) Mindfulness: Theoretical foundations and evidence for its salutary effects. *Psychological Inquiry*, 18(4): 211–237.
- Cary P (2007) A brief history of the concept of free will: Issues that are and are not germane to legal reasoning. *Behavioral Sciences and the Law*, 25: 165–181.
- Chesney J E (1939) Concept of *mens rea* in the criminal law. *Journal of Criminal Law and Criminology*, 29(5): 627–644.
- Dennett D (1986) Julian Janes's software archaeology. *Canadian Psychology*, 27: 149–154.
- Dennett D (2003) *Freedom Evolves*. New York: Viking.
- Eagleman D (2011) *Incognito: The secret lives of the brain*. Edinburgh: Canongate Books.
- Gallagher S (2005) *How the body shapes the mind*. New York: Oxford University Press.
- Libet B (2004) *Mind time: The temporal factor in consciousness*. Cambridge: Harvard University Press.
- Libet B (2011) Do we have free will? In *Conscious will and responsibility: A tribute to Benjamin Libet*, eds W Sinnott-Armstrong, L Nadel, 1–10. New York: Oxford University Press.
- Mele A (2009) *Effective intentions: The power of the conscious will*. New York: Oxford University Press.
- Pandey M, Lundrigan S, Moore C (2016) Excuses, justifications and blame in the offence narratives of convicted rapists in India. In the proceedings of the 72nd Annual Conference of American Society of Criminology, New Orleans, Louisiana.
- Rosch E, Mervis C B (1975) Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7: 573–605.
- Sorabji R (2004) The concept of the will from Plato to Maximus the Confessor. In *The will and human action: From antiquity to present day*, eds T Pink, M W F Stone. New York: Routledge.
- Strawson G (1994) The impossibility of moral responsibility. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 75(1/2): 5–24.
- Yaffe G (2011) Libet and the criminal law's voluntary act requirement. In *Conscious will and responsibility: A tribute to Benjamin Libet*, eds W Sinnott-Armstrong, L Nadel, 189–203. New York: Oxford University Press.
- Wegner D M (2002) *The illusion of conscious will*. Cambridge, MA: Massachusetts Institute of Technology Press.

RETRIBUTIVIST THEORY OF PUNISHMENT: SOME COMMENTS

Adebayo Aina

Department of Philosophy
Olabisi Onabanjo University
P.M.B.2002 Ago-Iwoye
e-mail: kofo@yahoo.com

Abstract:

The Retributivist approach to punishment attempts to address the challenges posed by utilitarian conception that punitive actions should strictly be associated with a cost-effective means to certain independently identifiable goods at the expense of justice. Justice proffers how the guilty deserve to be punished and no moral consideration relevant to punishment outweighs an offender's criminal desert. However, this just desert provokes difficulty in discerning proportionality between the moral gravity of each offence and the specific penalties attached. This consequently degenerates to another form of 'lex talionis' (revenge) in punitive justice.

Key words: Desert, justice, punishment, proportionality, retributivist, utilitarian.

Introduction

The following paper examines the retributivist theory of punishment. This theory attempts a response to some inadequacies proffered by the utilitarian principle that a right can only be cancelled on the ground of the expense of justice for a far greater number of people. Justice demands that the guilty be punished, in as much as they have unfairly appropriated benefits by failing to refrain from acting on inclinations which are deemed contrary to the way in which other members of society act. Also, burdens have been unfairly assumed by both the criminal and law-abiding citizens. So, in punishing the offender the misappropriated benefit is withdrawn by the imposition of a burden, thereby restoring an equita-

ble distribution of benefits and burdens which existed prior to the offence.

We shall first briefly discuss the meaning and nature punishment. Afterwards, we explore the basic features of the retributive theory of punishment, emphasizing that a morally tolerable account of punishment must include the idea that an offender, in committing an offence, deserves to be punished. We then proceed to consider some critique and responses beyond the theory.

A brief Discussion of the Meaning and Nature of Punishment

The general impression is that the concept of punishment involves the deliberate infliction of some kind of pain on an offender by a person or body of persons claiming the authority to do so. Thus, deserved punishment is conceived to "right the wrong of the criminal's disproportionate benefits-to-burdens ratio by imposing greater-than-normal burdens upon her" (Garcia, 1989, p. 270). That is to say "when I do something bad, I can lose or forfeit some of my normal moral rights against some unwelcome forms of treatment. Similarly, when I do something good, I may add to my rights certain special rights to form of treatment to which I am not normally entitle" (Garcia, 1989, pp. 263- 264). This calls for holding to account those, and only those, who properly are accountable when something unpleasant occurs. Against this submission, Antony Flew baldly states (though being descriptive and precise) five essential and defining characteristics of punishment: (1) it must involve the imposition of suffering (or deprivation of privilege); (2) it must

involve the suffering or deprivation of the offender; (3) the offender must suffer for his offence; (4) the suffering or deprivation must be imposed by a human agency; (5) the punishing agency must be authorized to impose the punishment (Loftsgordon, 1966, p. 342).

Implicit in the above characteristics are the form of hard treatment in punishment. These include: authoritative disavowal of the crime committed; symbolic non-acquiescence in it; vindication of the law which has been broken; and, placing blame squarely on the culprit, thereby absolving anyone else who might have been suspected (Primoratz, 1989, p. 188).

John Rawls contends that:

A person is said to suffer punishment whenever he is legally deprived of some of the normal rights of a citizen on the ground that he has violated a rule of law, the violation having been established by trial according to the due process of law, provided that the deprivation is carried out by the recognized legal authorities of the state, that the rule of law clearly specifies both the offense and the attached penalty, that the courts construe statutes strictly, and that the statutes was on the book prior to the time of the offense (Rawls, 1955, p. 8).¹

We may infer from these arguments that a just punishment must not only involve intentional deprivation of a person's normally recognized rights by the authority, using coercive means if necessary, but must also emphasize how the good it will achieve is so great that it outweighs the evil of the injustice involved (McCloskey, 1969, pp. 91–92). Indeed, it is necessary in a society which takes justice seriously to prefer the infliction of pain as a means of condemning harmful conduct which is prohibited by the law. To do otherwise would be a failure to protect and vindicate the rights of individuals over and against acts of misconduct performed by criminals.

According to Rawls, however, these measures presuppose that several conditions embedded in pure procedural justice have to be satisfied. First, the

offender must have been capable of not performing the act to which the liability is attached. Their liability for punishment is determined by their own acts and omissions as specified by laws. Second, the person must have performed the act with the knowledge of those implications which are attached to the crime. This is to say that they must have been conscious of the liability to suffer punishment as a necessary consequence for such an act. Still, the punitive severity must accord with the principle of proportionality: the more grave the crime, the more severe the deserved punishment. In other words, a punishment in this line of thought must not be so severe as to be either inhumane or “cruel and unusual.” Also, the rights of offenders, accused and convicted, must not be violated all for the sake of punishment.

Finally, the undertaking of a liability to suffer punishment was a necessary consequence of such an act. That is, unlawful harmful conduct is preferably prevented before the fact rather than punished after the fact. It is important that compliance under threat is much preferred to non-compliance followed by arrest, trial, conviction, sentence, and punishment. As a result, the knowledge of the lawful proscription is never retroactive in nature. It is legally appreciable if the offender consents to assume liability to suffer punishments involved in the voluntary commission of an offence with the knowledge that the liability is a necessary consequence of it (Rawls, 1955, pp. 1–13). Nino (1983) adds:

Unless we rely on the moral autonomy of the individual, making his liability to punishment depend on his free and conscious undertaking of it, all the burdens imposed on offenders, even in the name of treatment would be unfair even if they are not accompanied by tangible countervailing benefits. (p. 306)

Nevertheless, the value of punishment varies and is determined by the value of what one has been done. Thus Kurt Baier (1983, pp. 248–255) agrees to the fact that deserved punishment is deprivation imposed on someone only if he/she has been properly found to satisfy the condition(s) for which the system of punishment approves; but, it will not be a deserved punishment where a mistake in judging that the offender did satisfy the stated rules has occurred. Baier comments that, if the accused is found innocent, then an imposition of deprivation after that can neither be punishment nor a miscarriage of justice but rather “naked, blatant, shameless, mistreat-

¹ John Rawls (1921–2002) looks at punishment from his “justice as fairness” principle which engenders governing the assignment of rights and duties, and which regulates the distribution of social and economic advantages. See especially Rawls, John (1985). *A Theory of Justice*. Oxford: Oxford University Press; and Rawls, John (1985). ‘Justice as fairness: political not Metaphysical’. *Philosophy and Public Affairs*, 14(3): 223–251.

ment or victimisation” (1983, p. 249). Baier used this to draw attention to the distinction between punishment and victimisation with a much quoted article by K.B. Armstrong, noting that victimisation is a mistake because it would include such extreme cases as those of an innocent person being found guilty, sentenced to death, and subsequently executed by a court that had meticulously observed formal court procedures (1983, p. 250).

However, this distinction between punishment and victimisation is rather a conditional case. If it is presupposed and known that the imposition is victimisation, then it cannot be punishment, hence, no one who knows or believes that it is victimization can consistently believe that it is punishment; so, those who are unaware of its being victimisation merely believe — doing so wrongly — that it is punishment until they discover that it is victimisation. Thus, the fact that someone discovers that an imposition is a miscarriage of justice does not show that it was or is not punishment (1983, p. 250). It is therefore an issue of who can know what a punishment is and when victimisation emerges.

The retributivist theory of punishment

The retributivist justification of punishment is grounded on two *a priori* norms, namely: that the guilty deserve to be punished and no moral consideration relevant to punishment outweighs the offender’s criminal desert; and an epistemological claim, that we know with reasonable certainty what the guilty deserve. Criminal laws are similar to moral rules in that they all state standards of behaviour. But these are standards of society, organised into a state with its legal order, presupposing the authority of the state and its legal order. They are authoritatively formulated and applied solely through formalized procedures within appropriate institutions. Primoratz (1989) adds:

Legislative institutions of the state pass criminal laws which determine some of our most important legal rights, and turn their infringements into crimes. It is then up to criminal courts to condemn actions that violate such law and infringe the rights determined by them. By this, punishment vindicates the law broken, reaffirms the rights violated, and demonstrates that its violation was indeed a crime. Thus if there are to be rights sanctioned by the criminal law, if some actions are to be crimes, if there is to be criminal law at all,

there must be punishment as well. Where there is no punishment there are no crimes, no criminal law, no rights determined and sanctioned by such law. (p. 197)

The nature of criminal desert is that the offender must be blameworthy and that an offender deserved punishment because they have offended some one or set of statutes, and so a punishment must relate to their wrongdoing. Unlike the utilitarian theory, retributive justice involves a situation whereby the offender does not require secondary consideration of their desert, nor does it entail unfairly sacrificing individuals for the common good. The state has a right to punish a person on the ground that they have committed an offence in the past, and never merely because the state thinks the punishment will do more overall good than harm in the future. In this vein, punishment is deserved in as much as it is rendered permissible by one’s misdeeds. It can fit the crime in so far as it is the crime which determines the extent to which one’s moral rights are reduced, and thus determines the extent and severity of the punishment which has been rendered permissible. Pincoffs (1983) deduces two points from this inference: “the only acceptable reason for punishing a man is that he has committed a crime;” and, “the only acceptable reason for punishing a man in a given manner and degree is that the punishment is “equal” to the crime for which he is punished” (p. 267).

Hart (1968) also explicates three ways by which punishment cum retribution must depend:

R1: A person may be punished if and only if he has voluntarily done something wrong; R2: The punishment must match or be equivalent to the wickedness of the offense; R3: The justification for punishing is that the return suffering for moral evil voluntarily done is itself just or morally good. (p. 231)

Hart succinctly wants to say that, as reflected in the criminal law, the minimum condition required for punishment is that the person to be punished should him- or herself have done what the law forbids, at least so far as outward conduct is concerned; even if the condition is strict, it is not enough to render them culpable for sanction where someone else erred. It is pertinent to note that to say that a person is legally responsible for some act or harm is to state that their connection with the offence or harm is sufficient for punishment, according to the law.

From Hart's R1 above, possession of rights generally entails having duties to honour the same rights of others. It seems to follow plausibly from this contractual agreement that, when these duties are not fulfilled, the rights cease to exist. If this is true, then we seem to have a way of showing factually that the offender has forfeited their right not to be made to suffer. An offender, in violating the rights of others, forfeits either those rights they violated or an equivalent set of rights. Since it is often impracticable to prevent those rights which have been violated, we are justified in inflicting suffering in the form of the perpetrator losing those same rights. Hence, an offender deserves punishment because there is intrinsic justification in the guilty suffering.

The retributivist believes that the breaking of a law by the culprit is merely approximate to taking advantage where others choose not to; and, as this is wrong in itself, it deserves to be punished. Angelo Corlett (1986) contends that:

Punishment is justified on the ground that the wrongdoing merits punishment. It is morally fitting that a person who does wrong should suffer in proportion to his wrongdoing. That a criminal should be punished follows from the guilt, and the severity of the appropriate punishment depends on the depravity of his act. (p. 266)

From the above viewpoint, as supported by Hart's R2, it is important to note that it is not wholly unjust to punish the offender in order to render a punishment compatible with justice; but, we need to be cautious if we are to render a punishment fitting the crime. It seems plausible to rank crime in terms of its gravity, and thus in the following way: Each law creates a particular sphere of non-interference in such a way that laws prohibiting theft create a sphere of non-interference with regard to our property; and laws prohibiting murder create a sphere of non-interference with regard to our lives.² These spheres of non-interference can be ranked in terms of value, whereby the sphere defined by criminal laws would

presumably be more valuable than the sphere defined by property laws. Thus, we can rank the benefits that result from particular laws. With this ranking in mind, crimes can be ranked in order of their gravity in such a way that it would be plausible to argue that benefits deemed more valuable by the criminal call for more severe punishments. In other words, the graver the crime committed, the more severe the punishment (Burgh, 1982, pp.203–204).

This argument further indicates the moral apportionment of justice which resembles the principle of justice in the distribution of benefits and burdens. The idea here is that, in committing an offence, we do not think that the offender deserves punishment exceeding their crime, other than a degree of punishment proportionate to the gravity of the offence committed. Hence, the retributivist R3 justifies the amount of punishment in particular cases. While violating the rights of others involves forfeiting one's own rights, it is clear that violating specific rights of others does not entail losing all of one's own rights: if individual A steals \$200 from individual B, individual B (or anyone else) does not gain the right to impose all and any conceivable harms upon individual A in return; nor does individual A thereby become submissive for any suffering to which the community then wants to place upon them. Similarly, just as an innocent person can complain if forced to make severe sacrifices for the benefit of others, so a guilty person may claim that violation of any rights beyond those forfeited or alienated in order to benefit others is an injustice. If I ask which rights are forfeited in violating the rights of others, it is plausible to answer just those rights that one violates: violation constitutes forfeiture. But one retains those rights which one has continued to respect in others. Since deprivation of those particular rights violated is often impracticable, I am justified in depriving a wrongdoer of some equivalent right or set of rights, in terms of some average or normal preference scale, much like the one used by the utilitarian when comparing and equating utilities. Goldman (1979) comments further on this:

It would be difficult for a wrongdoer to complain of injustice when we treat him in a way equivalent to the way in which he treated his victim, provided that we also have a good ... reason for imposing upon him in that way. If he cannot demonstrate a morally relevant difference between himself and his victim, then he cannot claim that he must enjoy all those rights that he was willing to violate. But if we

² We need to note that the law of democratic countries requires that the person liable to be punished should, at the time of their crime, have had the capacity to understand what are required of them by law to do or not to do, to deliberate and to decide what to do, and to control their conduct in the light of such decisions. Other works that touch on this theme include Appadorai, A. (2005). *A Substance of Politics*. New Delhi: Oxford University Press.

deprive him not only of these or equivalent rights, but of one's far more important, whose loss results in far greater harm, then we begin to look like serious wrongdoer ourselves in multiplying violations of rights. It is at this point that the claim that two wrongs do not make a right begins to apply. A claim of injustice or victimization by the community made by the criminal begins to have merit, although in our anger at his wrongdoing, we are often unwilling to hear it. (pp. 45–46)

Given the viewpoint above, punishing an offender to a degree which exceeds what they deserve entails that we have treated them unjustly. In other words, any punishment in excess of what the offender deserves should be considered as imposing punitive sanction on an innocent individual. When we think in terms of forfeiting those rights one violates, or an equivalent set, there is no special difficulty there: one right or set of rights is equivalent to another for these purposes when an average preference scale registers indifference between the loss of either the one or the other. Justice then, from the retributive viewpoint, requires a principle of desert coupled with a principle of proportionality between the gravity of the offence and the punishment deserved.

By desert we mean a situation whereby "in committing an offence the offender forfeits his right not to be made to suffer" (Burgh, 1982, p. 197). This is to say that, by violating the rights of others in his criminal activities, the offender has lost or forfeited any legitimate demand that others honour all of their formally held rights. It incorporates the idea that distribution of benefits and burdens is just in such a way that everybody gets what he or she deserves. Hence the standard of desert first entails that people deserve punishment in proportion to the harm that their actions have caused. It leaves out entirely any considerations of blameworthiness, as accidental harm would be treated on a par with intentional harms. Also, people must be proportionally punished in accordance with a standard of moral iniquity. Finally, people deserve punishment in proportion to the degree to which they have upset the balance of justice. For example, it will be out of place to sanction a burglar for twenty years in jail while a proven murderer, beyond any reasonable doubt, is compelled to serve merely five years. This standard seems to presuppose adherence to a legal system which establishes a fair distribution of benefits and

burdens;³ the justification of punishment is then dependent on the justice of the society in question. This is as it should be, since we would not be justified in punishing people for violating laws that were unjust. At this juncture, in committing an offence, the offender deserves punishment. Herbert Morris (1968) puts the argument thus:

It is just to punish those who have violated the rules and caused the unfair distribution of benefits and burdens. A person who violates the rules has something others do not have—the benefits of the system—by renouncing what others have assumed, the burdens of self-restraints, he has acquired an unfair advantage. Matters are not even until this advantage is in some way erased. Another way of putting it is that he owes something to others, for he has something that does not rightfully belong to him. Justice—that is, punishing such individuals—restores the equilibrium of benefits and burdens by taking from the individual what he owes, that is, exacting the debt. (p.478)

The idea is that in punishing the offender, benefits must be taken away as a form of desert, replaced in turn with a burden "thereby restoring an equitable distribution of benefits and burdens which existed prior to the offence" (Burgh, 1982, p. 203). This injustice consists in the wrongdoer's having received the benefits of others' self-restraint in addition with benefits of their own lack of restraint. Morris's position is straight forward: a person gains benefits from the self-denial of others. In sum, the wrongdoer receives the benefit of others' self-denial and unfairly gains the additional benefit of failing to shoulder the burden of their own self-denial towards others.

Retributivist theory involves the re-distribution of rights after some initial distribution of them has been redistributed by the offender. Here, if we are to morally justify punishment of offenders, we must proffer the reasons for their forfeiture of rights. We may argue that they have lost the earlier legitimate demand that others honour their rights by virtue of violating the rights of others through their criminal

³ It is this standard that Rawls aims at in his original position: regarded as the embodiment of the various substantive normative claims at the heart of "justice as fairness," the principle is intended to model the idea that society should be regarded as a fair scheme of cooperation between citizens conceived as free and equal. See Rawls, John (1985). *A Theory of Justice*. Oxford: Oxford University Press.

activities, gaining unfair benefits in the process (Goldman, 1979, p. 43). Examples of unfair benefits could be: (1) ill-gotten gain, such as money from robbery or cyber-crime; (2) not bearing the burden of self-restraint, as expected in the implicit contractual agreement; (3) pleasure derived from committing a crime; and (4) the sphere of non-interference which results from general obedience to the universal law violated. So, when the criminal is punished retributively, “right reasserts itself by negating this negation of itself” (Day, 1978, p. 505).

Retributive justice then requires that wrongdoers receive no more and no less than what is proportionate (or just) to their crimes. People deserve to be treated in the same way that they voluntarily choose to treat others. However, there is a dangerous tendency to slip from retributive justice to an emphasis on revenge. Scholars like Nozick have critiqued retributivism, legal or moral, on the ground that lack of tacit or express appeal leaves it little distinct from mere retaliation or revenge (Nozick, 1981, p. 366–74). However, we are not comfortable with this scholarly submission on the premise that retributive justice distinguishes itself from *lex talionis* (revenge) or vengeance in several relevant ways. First, whereas revenge strikes out at real or perceived injury, retribution speaks to an objective wrong. Secondly, whereas revenge is wild and insatiable, retribution acknowledges the moral repugnance of assigning unjust punishment to petty crime as well as assigning light punishment to violent, heinous crime. Also, vengeance is characterised by injury and delight in bringing evil upon the offending party, but retribution has as its weapon a greater social good and takes no pleasure in punishment. Finally, the avenger aspires for retaliatory mood due to something done to them or to their group, whereas retribution is impersonal and therefore demands impartiality, not being subject to personal bias (Maiese, 2004).

To this end, retributive theory implies that any punishment of a person by the state is morally justifiable, if and only if they have done something which is both a legal and a moral offence, and only if the penalty is proportionate to the moral gravity of his offence. This seems satisfactory, but it does not go far enough. As such, we shall next consider some of the implications of the principle of proportionality.

Beyond the Retributive theory of punishment

Justice demands that the principle of desert must be within the confines of the balance between

the crime and the degree of punishment deserved. But it is difficult to discern such equivalence between the gravity of each offence and specific penalties attached. This is so because the terms of the equation are not commensurable. Besides, we need to note that every excess over the just amount must be in the same ethical position as punishment of the innocent, an injustice which seems much worse than non-punishment of the guilty. So, the notion of desert altogether, such as the ranking of crimes and punishments, may take retributive punishment to institutionalise a vendetta mentality which may be offensive and dehumanising.

It is not impossible that retributive justice may allude to one or another form of *lex talionis*, as this precludes the sufficient proportionality principle in society. For example, there are some problems with the particular distribution of burdens and benefits of punishment. It is incoherent, frequently inapplicable, and sometimes repugnant to evaluate and determine the punishment, with respect to the principle of proportionality, to be meted out on rapists, speeders, drunk drivers, embezzlers, and forgers, other than the serving of jail terms. In other words, it will be morally unjust to rape a rapist in order to certify the principle of proportionality. It portends no good argument for retributive punishment other than a vindictive mentality that is offensive and degrading to human beings in that any excessive punitive measure against the offender is objectionable for its imposition of an equivalent amount of punishment on an innocent person, as being canvassed by and against the utilitarian doctrine (Philips, 1986, p. 379–8). About the best the notion of fittingness can do is to suggest a rough correlation between the severity of the crime and the severity of the penalty. Hence, there is then the need to evolve a mixed theory wherein this distinctive proportionality principle on the concept of punishment could be resolved. Miller’s reformatory theory of punishment is appropriate here.

Miller (1970) refers to the traditional theories of punishments as the *esse* and the *bene esse*, retributive and utilitarian respectively. Miller is worried about the underlying effectiveness of the above theories to initiate *Plene esse* as a justification of punishment akin to prescriptive language and aimed at influencing behaviour. *Plene esse* aims to show that an individual does have a right to determine their own style of existence, but not in respect of deterrence or reform in terms of retaliation; rather, it is to emphasize that society also has the right to express

disapproval, and — if necessary — attempt to restrain such a life style. Indeed, a good action is that which is willed by “means of the conception of reason” which is acceptable and “valid for every rational being as such” (Kant, 1785/1990, p. 256). It is a form of an apodictic principle enhanced by reason itself. This represents an end sound for every rational being, so long as individual’s reflection is free from the attraction of desire and inclination. In other words, *Plene esse* requires that the views of society should not only be pronounced, but their recognition must be accorded to the individual — who may not share the views of the society — with the intent that the individual works for a change of attitudes within the confines of the societal laws. In turn, society can work for a change of attitudes within that of the individual. Thus, the punishment of those who offend against the laws of society is a means of expressing this disapproval. Such unpleasant experiences may restrain an individual from further pursuit of their lifestyle, or the threat of it may even deter them from adopting it in the first place (Miller, 1966, pp. 256–260).

This will influence behaviour rather than changing it. However, Miller misses the point: his position only attempts to address the effectiveness of the punishment based on the conviction that individuals are responsible agents. If individuals are responsible moral agents, then there would be no need for authorised agents to not only monitor human actions but also to enforce the consensual laws of society. Also, Miller believes that the penal desert of this nature, between wrongdoing and censure, would have produced pain to an offender for their offence. Its application within the public sphere through criminal justice systems would merit a situation whereby the person censured accepts the censure as justified, and they would thus be motivated to avoid crime in the future. Our objection to this submission is that reflective apologetic reparation must be voluntary if it is to be of any value, instead of being expressive or communicative in character by the community concerned. As a result, it cannot resolve the problem of ineffectiveness in the control of a society in which the traditional theories of punishment are culpable. Even legitimate rights to morally punish in this reformatory theory may flourish in as much as it is consensually agreed to by all rational beings in society, but it remains short-lived within the relation between the ideal and the actual.

It becomes imperative for us to reconsider the preconditions which sometimes influence behaviour

in society, consequently punishment. We agree with Odera Oruka’s submission in this respect. An abolitionist, Odera Oruka examines the implications of the concept of pre-conditions of just punishment in his *Punishment and Terrorism in Africa*. He denounces the importance of both utilitarian and retributivist conceptions of punishment with the view of upholding the empirical foundations of the criminal mind (Oruka, 1972, p. 2). These foundations revolve around the economic imbalance or condition which a person finds him- or herself difficult to control. He contends that “[criminal] factors have to do with such things as irresponsible parental care, belonging to a despised or poverty-stricken class, discrimination and suppression by the family or society, being amoral or [a] social outcast, mental derangement, a bad education” (p.18).

Nevertheless, in order to abate these forces that engender crimes, he suggests a balance between social security and social harmony. He opines that reason and not force of will serves as the means through which social values, social harmony, and cooperation are primal, to the extent that society is free of tensions and security is consequently brought to the fore (Oruka, 1972, p. 29). He concludes with the recommendation of social treatment for criminals (Oruka, 1972, p. 92), premised on the fact that “crime is an illness requiring treatment by psychiatrists and psychologists” (Menninger, 1983, p. 255).

Conclusion

We have been able to show the strengths and weaknesses of the retributive theory of punishment, where the retributivist view is such that the state has the right to punish any culprit who contravenes the law — but it is not because the state is convinced that the punishment will do more overall good than harm in the future. Thus society is entitled to impose penalties on criminal acts, in as much as they are immoral and that the only legitimate function of punishment is to exact just deserts on the perpetrators of immoral deeds. Unfortunately, the retributivist claim that the resentment or indignation directed toward offenders is fitting, rather than merely ill-disguised anger, is based upon the assumption that the criminal laws, whose violation makes one eligible for punishment, protect individual rights.⁴ The

⁴ In addition to what was alluded to earlier in the paper, the rule of law, which mirrors moral duties, has (among other things) an educational function. It is formulated in a way which, though not infallible, is in some sense

defects of this conceptual approach revolves around the desert offenders ought to satisfy or experience, for it is noted that retributive justice may be driven by one or another form of *lex talionis*. Nevertheless, Miller's mixed theory conception is incapable of addressing the challenges. However, Oruka's insistence for the concept of preconditions for just punishment would meet the challenges as criminal behaviour is more of an illness than the deployment of some punitive sanction by the state. Of course, this in-itself may be counter-intuitive, in as much as crime in-itself is grounded on the incontinent nature of a person. Any attempted excuse for sanctions may thus produce unwholesome outputs for the society in the long run.

References

- Baier K (1983) The concept of punishment and responsibility. In *Morality in Practice*, ed. J P Sterba. Belmont, California: Wadsworth Inc.
- Burgh R W (1982) Do the guilty deserve punishment? *The Journal of Philosophy*, 79(4): 193–210.
- Corlett J A (1986) Making sense of Retributivism. *Philosophy*, 76(295): 77–110.
- Day J P (1978) Retributive punishment. *Mind: New Series*, 348: 498–516.
- Garcia J L A (1989) Deserved punishment. *Law and Philosophy*, 1(8): 263–277.
- Goldman A H (1979) The paradox of Punishment. *Philosophy and Public Affairs*, 9(1): 42–58.
- Hart H L A (1968) Prolegomenon to the principles of punishment. In *Philosophy, Politics and Society*, eds P Laslett, W G Runciman. Second Series. Oxford: Oxford University Press.
- Hart H L A (1975) Responsibility. In *Responsibility: Selected Readings*, eds J Feinberg, H Gross. California: Dickenson Publishing.
- Kant I (1785/1990) Fundamental Principles of the Metaphysics of Morals. *The Britannica Great Books*, 42, 2nd edition. Chicago: Encyclopedia Britannica.
- Lofts-Gordon D (1966) Present-day British philosophers on punishment. *The Journal of Philosophy*, 63(12): 341–353.
- Maiese M (2004) Retributive justice. In *Beyond Intractability*, eds G Burgess, H Burgess. Conflict Research Consortium, University of Colorado, Boulder.
- Mccloskey H J (1969) Utilitarian and retributive Punishment. *The Journal of Philosophy*, 64(3): 91–110.
- Menninger K (1983). The crime of punishment. In *Morality in Practice*, ed. J P Sterba. Belmont, California: Wadsworth Inc.
- Miller W A (1970) A theory of punishment. *Philosophy*, 45: 307–316.
- Miller W A (1996) Mr. Quinton on "An odd sort of right." *Philosophy*, XLI: 258–260.
- Morris H (1968) Persons and punishment. *Monist*, LII(4): 475–501.
- Nino C S (1983) A consensual theory of punishment'. *Philosophy and Public Affairs*, 12(4): 289–306.
- Nozick R (1981) *Philosophical Explanations*. Cambridge, MA: Harvard University Press.
- Oruka H O (1976) *Punishment and terrorism in Africa: Problems in the philosophy and practice of punishment*. Kampala: East Africa Literature Bureau
- Pincoffs E L (1983) Classical retributivism. In *Morality in Practice*, ed. J P Sterba. Belmont, California: Wadsworth Inc.
- Primoratz I (1989) Punishment as language. *Philosophy*, 64(248).
- Rawls J (1955) Two concepts of rules. *The Philosophical Review*, 64(3), 1–13.

authoritative, and thus the content of the shared morality. See Mclean, G.F and Ellrod, F.E. Eds. (1992). *Philosophical Foundations for Moral Education and Character Development: Act and Agent*. Washington D.C.: C.R.V.P.