2024-05-17

The Criminal Brain: Neurointerventions and Mental Freedom

Craig, Jared N.

UNIVERSITY OF CALGARY

The Criminal Brain: Neurointerventions and Mental Freedom

by

Jared N. Craig

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PHILOSOPHY

CALGARY, ALBERTA

MAY, 2024

Abstract

This dissertation addresses the ethics of non-consensual neurointerventions for criminal offenders by considering 'punishment equivalence arguments.' They suggest if proven safe and effective, neurointerventions could serve as ethically viable alternatives to traditional punitive measures like imprisonment. I examine the four premises of these arguments: (1) the state's legitimate authority to punish, (2) the assumed safety and efficacy of neurointerventions, (3) the similar effects of direct and indirect interventions into the brain, and (4) that based on specific ethical paradigms, neurointerventions are equivalent to standard punitive measures. Practically, I highlight challenges posed by the current crisis in our penal systems and the nascent state of our understanding of neuroscience. Theoretically, I argue unique properties of neurointerventions pose distinct threats to mental freedom. The project supports a negative claim: punishment equivalence arguments fail to offer meaningful guidance in a broad range of foreseeable cases or mitigate deep-seated ethical reservations. However, it also suggests a positive claim: there are compelling moral and prudential reasons for extreme caution before rushing to implement a comprehensive project for neurointerventions in criminal justice practices. There is a pressing need for further ethical theorizing, including the possibility of recognizing rights over the brain and mind.

Keywords Criminal sentencing • Neurointerventions • Human rights • Incarceration • Mental integrity • Parity Principle • Neuroethics • Freedom of Thought

Preface

This dissertation incorporates revised sections of a previously published paper titled “Incarceration, Direct Brain Intervention, and the Right to Mental Integrity – A Reply to Thomas Douglas” in *Neuroethics* (2016) 9:107–118.

The remaining content of this dissertation is original and unpublished, representing the independent work of the author. The author expresses deep gratitude to the numerous individuals who provided valuable discussions and assistance throughout the research process, with special mention in the acknowledgements section.[1]

Acknowledgements

A significant part of this dissertation examines how our relationships with others shape the brain, the mind, and our understanding of meaning and place in the world. The research journey spanning over a decade has been transformative, instilling in me a profound sense of humility and gratitude for the contributions of those around me.

Throughout this journey, I have become acutely aware of my limitations and grown to recognize that this dissertation is not solely the result of my own efforts. Rather, it is the culmination of inspiration, steadfast support, invaluable assistance, and extraordinary patience from numerous individuals who have played an indispensable role in its creation. I seize this moment to extend my sincere appreciation for their contributions.

First, my deep gratitude goes to my supervisor, Allen Habib, for his patience, insightful guidance and unwavering encouragement, demonstrating a remarkable commitment to fostering my academic development. I extend my deepest gratitude to my dissertation committee members—Greg Hagen and Ann Levey. Despite their busy schedules, they provided invaluable assistance, significantly enriching my understanding of the field and improving the quality of this work. I am profoundly appreciative of their dedication, which has been instrumental in shaping the final outcome. I am sincerely thankful for their unwavering support and guidance.

Additionally, I am grateful to the external examiners, Ted McCoy and Jennifer Chandler, for their thorough reviews and constructive critiques which have greatly improved this dissertation. Their expertise and insightful comments have been exceedingly beneficial.

I am deeply indebted to Walter Glannon, my previous supervisor, whose contributions during the initial stages of my research were invaluable. His remarkable expertise and groundbreaking work in the field of neuroethics not only shaped the discipline itself but also ignited my passion and dedication to delve deeper into this subject. His guidance has played a pivotal role in shaping the trajectory of my studies, and I am immensely grateful for his support and mentorship.

I am also grateful for the substantial funding support provided by the Social Sciences and Humanities Research Council. Their commitment to advancing scholarly excellence has significantly aided my academic journey. Their financial support highlights the importance of investing in the intersecting fields of philosophy and neuroscience.

Denise Retzlaff deserves my profound appreciation for her consistent support throughout my ten-year journey within the philosophy department at the University of Calgary. Equally, I’m thankful

to the Department Heads and Graduate Program Directors, including Ishtiyaque Haji, Marc Ereshefsky, Richard Zach, and Nicole Wyatt, for their faith in my abilities and their numerous accommodations.

Over the years, I've had the pleasure of forming supportive and inspiring friendships among my colleagues. Aaron Thomas-Bolduc, Justine Caouette, Walter Reid, Myrium Bronski, and Robert Armstrong, your contributions have truly enriched this journey. Remembering our colleague and my close friend, Sinan Sencan, I hope he finds peace in his rest. I am equally grateful to my students, whose inquisitiveness and enthusiasm have been a constant source of inspiration. Their unique perspectives have enriched my understanding and contributed significantly to my personal and professional growth.

I would like to extend my heartfelt gratitude to Andrew Craig for his unwavering support during challenging times and for providing invaluable insights—through word and example—and whose wisdom greatly influenced the work presented in this dissertation. Your guidance and encouragement have been instrumental throughout this journey. I would also like to express my sincere appreciation to Terry Penner for his invaluable assistance, especially during the final stages of completing this dissertation.

In embarking on my journey into philosophy, I am immensely indebted to the guidance and support of Dr. Howard Hopkins and the late Ernie McCullough. These esteemed philosophers not only possess remarkable intellect and expertise in the field but also cherished friends whom I hold in the highest regard. Their profound insights and unwavering support have shaped my understanding and enriched my philosophical exploration. I will always cherish the memories and the friendship we shared. I owe a significant part of my professional path to Glenn Luther, my first-year criminal law professor. His commitment to criminal defence ignited my passion and paved the way for my career.

Equally, I am profoundly grateful to my law practice partners, Gilliana Shiskin and John Hooker. Their steadfast support, camaraderie, understanding, and personal sacrifices—notably during my absence from the office while completing my studies—as well as their accommodations along this journey, have been invaluable. Their unyielding faith in my potential has been a constant source of motivation, and I deeply thank them for their companionship on this journey.

Throughout John, Jill, and I's process of building our law practice and serving our clients, we have been incredibly fortunate to receive invaluable assistance and insights from an exceptional group of individuals. Their support and guidance have played an integral role in our professional growth and success. I would like to express my sincere gratitude to the dedicated support staff, law students, and articling students who have played a significant role in this endeavour over the past years. Among

them, I would like to extend a special appreciation to Shervin Sabat, Marshall Wolff, Brian Lines, Leanndria Halcro, Gunntas Sidhu, Corey Wutzke, Spencer Craig and McKinley Pilling, among many others. Their unwavering commitment, hard work, and exceptional contributions have made a lasting impact on both our clients and the development of this work. I am truly grateful for their invaluable support, and I deeply appreciate their significant contributions.

To the clients I've had the privilege of representing during my tenure as a criminal defence lawyer, I express my heartfelt thanks. Your bravery and teachings about resilience have illuminated the remarkable strength of the human spirit and its ability to uncover glimmers of hope amidst adversity and in the darkest of times. Special recognition is owed to T.L., a young man of promise and courage whose life ended prematurely, as well as other clients whose absence is deeply felt. The lives and memories of these individuals have served as powerful reminders of the utmost significance and inherent value of every member of the human family and the urgent necessity for reforms in our legal practices. You have been a deep well of inspiration, guiding me through personal challenges in completing this dissertation.

I would like to express my heartfelt appreciation to my exceptional parents, John and Carol. Your boundless love, unwavering support, and exemplary guidance have been instrumental in shaping the person I am today. You have always been there for me, providing valuable life lessons that have instilled in me the values of dedication, hard work, compassion, and the significance of family. I am forever indebted to you for the countless ways you have enriched my life. Throughout the years, I have accumulated a debt of gratitude that I know I will never fully repay.

To Lisa, my incredible wife, your role in the journey of writing this dissertation has been nothing short of vital. Over the past six years, as I navigated the complexities and challenges of this academic pursuit, your unwavering support and inspiration have been my guiding light. Through every hurdle and each forward leap, you have been my steadfast companion, sharing in every instant and motivating me to continue steadfastly and aim for greater heights. Your belief in me has been a powerful force, helping me to become a better version of myself both academically and personally. As I approach the final stages of completing this PhD, it is with the strength fostered by your love and support. It stands as much as a testament to your unwavering presence in my life as it does to my own dedication. Lisa, my boundless gratitude for you is interwoven into every page of this work.

As I conclude, it is with profound gratitude that I dedicate this work to my four exceptional children—Reiner, Hannah, Jordan, and Jace. Your presence brings constant joy, motivation, and purpose into my life, and your sacrifices throughout the lengthy period of my studies, which began when you were still very young, have not gone unnoticed.

To Jace, my youngest, your bright curiosity and delightful sense of mischief are qualities I am confident will serve you well throughout life's adventures. Jordan, your kindness and bravery continue to inspire me each day. Hannah, your wisdom far exceeds your years, and your innate curiosity perpetually astounds me. Reiner, your patience, unwavering compassion and courageous nature have been a constant source of inspiration for me. I am deeply grateful for the valuable lessons you have taught me throughout our journey together. To each of you, your influence on my life is immeasurable, and I will forever cherish the profound impact you, my children, have had on shaping the person I am today. Thank you for being an exceptional teacher and an incredible source of strength.

Dedication

For Reiner, Hannah, Jordan, and Jace.

Table of Contents

Introduction

Neurointerventions and Crime Prevention – Punishment Equivalence Arguments

In *The Shawshank Redemption*, a renowned short story, screenplay, and film, Andy Dufresne is serving a sentence of life imprisonment for a crime he did not commit (Darabont & King, 1994b; Darabont & King, 1996; King, 2010). Aiming to liberate his fellow inmates from the realities of prison life, he infiltrates the prison warden's office. He plays *Mozart's The Marriage of Figaro: Duetto-Sul Aria* through the prison yard loudspeaker system. Recounting his two weeks of solitary confinement in 'the hole,' Dufresne explains:

> easiest time I ever did… the music was here [gesturing to his head] … they can't take that from you. We need it, so we don't forget… That there are things in this world not carved out of gray stone, that there's something inside that they can't get to, that they can't touch (Darabont & King, 1994b).

The Shawshank Redemption is often cited as an illustration of the devastating effects of the prison environment—the concept of 'institutionalization' or 'prisonization' (Gillespie, 2002; Weizhi, 2010). But the protagonist's liberation from Shawshank also symbolizes a realm of inner liberty. One that is perceived as untouched by external influences and beyond the state's grasp. A form of 'mental freedom,' standing as a testament to the profound strength of mental self-determination (Craig, 2016, p. 107).

Contemporary Advancements in Neuroscience—The Criminal Brain and Neurointerventions

During the 21st century, advancements in neuroscience have made remarkable strides in unravelling the intricate workings of the human brain, illuminating how a three-pound organic mass has played a key role in the emergence of consciousness, rationality, morality, and even the

fabric of human societies. Alongside a growing understanding of the intricate workings of the brain, research has also yielded a powerful arsenal of tools for brain intervention known as 'neurointerventions'—cutting-edge technologies such as novel psychopharmacological agents, advanced neuromodulation techniques, neuroprosthetics, brain-computer interfaces, and optogenetics.

In contemporary discourse, these advancements have prompted discussions about leveraging such technologies to target the 'criminal brain'—neurobiological mechanisms linked to criminal or antisocial behaviour. This has spurred divisive debate in the field of neuroethics about ethical issues surrounding the use of neurointerventions as an alternative to traditional punitive measures, such as incarceration, or to assist in offender rehabilitation.

The Emerging Debate

The potential of forcible intervention in the human brain echoes with visions immortalized in dystopian fiction (Burgess, 1962; Forman, 1970; Huxley, 2021 [1932]; Orwell, 1977 [1949]). But these issues are far from novel: "[f]antasies of intervening in the psyche and the use of psychotropic agents have a venerable history in mankind" (Merkel et al., 2007, p. 11). Surely, our brains and minds change daily through traditional means—with traditional punitive measures often designed to achieve precisely this. Thus, if we assume these technologies are safe and effective, one might argue there is nothing fundamentally different about them. At very least, we should rigorously scrutinize our visceral intuitions and determine whether they withstand rational scrutiny.

To leave no doubt, crime is a pressing social issue. As Kant notes, the presence of "evil" can be traced across all human civilizations (Kant, 1998); throughout human history, Hegel vividly characterized as a "slaughter bench" (Hegel, 1956, p. 35). There are those among us capable of committing heinous and abhorrent acts that visit severe and often irreversible harm to others and warrant our unequivocal condemnation. Further, in some cases, criminal behaviour is tied to severe cognitive and emotional impairments and underlying neurobiological dysfunctions.

It is widely accepted that our biological nature contributes to some aspects of evil in the world.[2] The Enlightenment era fostered a belief in humanity's ability to transcend natural limitations, leading to groundbreaking advancements in medicine and biomedical research.

Notably, the discovery of antisepsis, tuberculosis treatments, sulfa drugs, and the polio vaccine showcases our capacity to reshape nature for the better (Bronstein, 2010, p. 85).

Given this context, the potential of novel neurotechnologies to address criminal or antisocial behaviour—a persistent malady that inflicts harm on our species—becomes a compelling prospect. These technologies could effectively mitigate the societal costs of crime and facilitate offender rehabilitation, potentially satisfying a broad spectrum of theoretical justifications for punishment—from retributive, consequentialist, and any number of objectives.

But transitioning from abstract theory to the tangible complexities of real-world applications, the intersection of medical science and criminal justice has historically been fraught with ethical dilemmas. Historically, a myriad of medical interventions, including forced sterilizations, lobotomies, electric shock therapy, and physical castration, have been employed under the guise of 'treating crime'(Arboleda-Florez, 2005; Buchanan, 2007a; Carlson, 2001; Faria Jr, 2013; Matravers, 2018, p. 74; Mayer, 1947; McTernan, 2018a, p. 274; Ryberg, 2020, p. 170; Silver, 2003; Spece, 1972). Some of these troubling practices were even scrutinized during the Nuremberg Doctors' Trial, as defence attorneys sought to establish moral comparisons with medical experiments conducted in other Western nations (Comfort, 2009; Hornblum, 1997). Such practices underscore the profound ethical quandaries inherent in melding medical intervention with penal objectives, casting a long shadow that lingers over the discourse surrounding contemporary neurointerventions (McTernan, 2018a, p. 274; Ryberg, 2020, pp. 167-168).

The rapid advancements in neurotechnology have brought us to the precipice of significant ethical dilemmas, particularly concerning neurointerventions and their impact on mental freedom. This critical juncture invites us to ponder the boundaries of liberty—exploring the untouched sanctum of the mind, reminiscent of the indomitable human spirit celebrated in literature. Yet, as we venture further into this territory, we find that our initial apprehensions have been largely overlooked rather than comprehensively addressed, marking a conspicuous void in ongoing discussions.

The complexity of these issues cannot be understated. With the advent of advanced neurointervention technologies and a deepening understanding of brain mechanics, scenarios once confined to dystopian narratives now edge closer to our reality. The shift from fiction to plausible future compels a more nuanced and rigorous exploration of the ethical landscapes that

govern the use of neurointerventions. Whether as alternatives to incarceration or as novel modalities for criminal rehabilitation, the conversation must evolve. It is time to expand our dialogue, confronting these challenges with the gravity and depth they demand.

Chapter Overview

This dissertation delves into the complex discussions surrounding compulsory neurointerventions on individuals convicted of crimes, specifically those administered without their consent (CNIs). It focuses on dissecting and critiquing Punishment Equivalence Arguments (PEAs), which propose that if presumed safe and effective, CNIs could be considered ethically on part with traditional forms of punishment such as incarceration (TPs).

PEAs frame a philosophical discourse that employs a structured analysis comprising a central question, evaluative metrics, and a sequence of propositions, along with their objections and responses. This dissertation unpacks these elements across four pivotal premises:

- **Justification of Punishment**: A fundamental examination of why punitive measures are necessary.
- **Conditional Efficacy and Idealization of CNIs**: The theoretical assumption of CNIs' safety and effectiveness.
- **The Parity Principle**: Evaluating the ethical significance of intervention methods.
- **Claim of Equivalence**: Comparing the impacts of CNIs with traditional punishments (TPs).

The opening chapter sets the stage for an in-depth exploration of the dissertation methodically:

1. **Outlining the Debate's Core Components:** Introducing the major themes and questions at the heart of the discussion.
2. **Formalizing PEAs:** Formalize PEAs, setting out their key premises and conclusions.
3. **Reviewing the Literature:** Surveying significant contributions and perspectives within the scholarly conversation.
4. **Establishing Key Terminology:** Defining essential terms and concepts critical for understanding and engaging with the arguments.
5. **Previewing the Following Chapters:** Offering a brief overview of each subsequent chapter, highlighting their focuses and connections to the overarching thesis.

Core Components of the Debate

This dissertation explores the ethical debate surrounding the use of compulsory neurointerventions (CNIs) as alternatives to traditional punishments (TPs). It outlines how Punishment Equivalence Arguments (PEAs) contribute to the wider discourse on criminal justice ethics. PEAs represent just one perspective within a multifaceted debate. The arguments are structured into premises and conclusions, streamlining the discussion to enhance clarity and facilitate a thorough examination of the complex issues involved. While this is a technical exercise, it is important to understand the formal components of the debate, and it will pay dividends in the long run.

The Motivating Question

The debate over using compulsory neurointerventions (CNIs) as alternatives to traditional punishments (TPs) within the criminal justice system hinges on a crucial question: If traditional punishments are justifiable, why not consider CNIs?

- "We send convicted criminals to prison without giving them another choice. If we can do that, why can we not require a direct brain intervention?" (Greely, 2012, p. 164);
- "If locking offenders in prison for a long period is justified, then it is difficult to see why requiring prisoners to undergo some type of safe and effective neurointervention couldn't also be acceptable" (Douglas, 2014a);
- Is it really plausible to suggest neurointerventions are 'always' objectionable, while "it is acceptable to lock up people behind bars for years against their will" (Ryberg, 2018, p. 180)?

In simpler terms, If we think it is okay to punish criminals through TPs, such as prison, why not consider CNIs as an alternative?

The Claim of Moral Objectionability

The *Claim of Moral Objectionability* is a primary response to the use of compulsory neurointerventions (CNIs) as punishment. This claim argues that CNIs are morally objectionable according to a specific ethical metric, M, which assesses the 'wrongness' of an action:

P1 (Premise 1): CNIs are morally objectionable based on metric M, which reflects a specific ethical standard.

P2 (Premise 2): The state should not engage in actions deemed objectionable by M.

C (Conclusion): Therefore, the state should refrain from employing CNIs.

In simpler terms, imposing neurointerventions is considered wrong as it contravenes fundamental ethical principles regarding the treatment of individuals. The definition of 'wrongness' varies across different moral, ethical, or legal theories and centers on Metric M—a measure of moral objectionability. As Ryberg notes, "the use of involuntary treatment will probably strike many as highly objectionable—in particular in light of the significance that is today attributed to autonomy and self-determination in medical health treatment" (Ryberg, 2018, p. 180).

Response: The Claim of Criminal Liability

The *Claim of Criminal Liability* counters the *Claim of Moral Objectionability*, suggesting criminal actions warrant responses that might bypass violate M:

P1 (Premise 1): Accept the *Claim of Moral Objectionability* based on M.

P2 (Premise 2): Committing a criminal act incurs a form of moral liability, leading to criminal liability, aligned with specific theories of punishment P (e.g., retributivism, rehabilitative measures), even when these measures raise ethical questions under M.

P3 (Premise 3): Given the justification for TP, under P, CNIs can be considered justified under similar circumstances.

Conclusion (C): Therefore, under specific conditions and theories of punishment, the use of CNIs is justified under theory P, despite initial objections.

This argument suggests that although CNIs may initially seem wrong for criminals, their use could be justified since criminals are subject to different treatment. It questions the notion that CNIs are invariably wrong in such contexts.

Objection: Safety and Efficacy

A further objection to the use of CNIs as punishment is concerns about safety and effectiveness:

P1 (Premise 1): Accept the *Claim of Moral Objectionability* based on M.

P2 (Premise 2): Accept the *Claim of Criminal Liability* based on P.

P3 (Premise 3): The ethical acceptability and justification of employing CNIs as punishment depends on their proven safety and effectiveness in treating criminal behaviour.

P4 (Premise 4): Current evidence does not sufficiently demonstrate CNIs' safety and efficacy, presenting unresolved risks and ethical concerns not inherently associated with TPs.

Conclusion (C): Therefore, CNIs currently fail to meet the ethical standards set by metric M for safety and effectiveness, undermining their justification as a comparable alternative to traditional punishments.

In simpler terms, the proposition of using CNIs for punishment is problematic due to uncertainties about their safety and effectiveness. As it is widely cautioned, with respect to the brain, a complex and vital organ, "new psychoactive medications and the direct stimulation of the brain may have unforeseen and terrible consequences" (Elfferich, 2021, p. 132; Greely et al., 2008). The lack of definitive evidence on the safety and efficacy of CNIs presents a significant ethical barrier to their implementation.

Response: Assumption of Safety and Efficacy

One response to this claim is that for the purpose of ethical theorizing, we can accept that neurointerventions may pose risks to health safety, but this need not *necessarily* be the case. If we assume neurointerventions are safe and effective, they remain justifiable under that condition:

P1 (Premise 1): Accept the *Claim of Moral Objectionability* based on M.

P2 (Premise 2): Accept the *Claim of Criminal Liability*

P3 (Premise 3): Accept the *Objection of Safety and Efficacy*

P4 (Premise 4): While the ethical acceptability of CNIs is questioned due to current evidence gaps in safety and efficacy, we can engage in a hypothetical exploration assuming these interventions are both safe and effective.

> Conclusion (C): Assuming CNIs are both safe and effective (P4), they might ethically align with theories of punishment P, despite objections based on metric M, given the moral implications of criminal acts (P2).

The idea here is to think about what would happen if CNIs were proven to be safe and effective in the future. This allows us to explore the potential of neurointerventions in the justice system without being limited by current technological shortcomings. This may allow us to dispel problematic intuitions, identify issues that are "genuinely urgent" (Ryberg, 2020, pp. 11-12), and facilitate "anticipatory ethics," which is proactive to the "rapidly evolving state of neuroscience itself" (Farah, 2011a, p. 776; Vincent, 2014, p. 21). In other words, by assuming CNIs work as intended, we can consider their ethical implications and potential benefits, paving the way for future discussions and developments in this field.

The Direct Brain Intervention (DBI) Objection

A further objection, the *DBI Objection* notes that CNIs, through direct brain intervention (DBI), raise unique ethical concerns not found with traditional punishments like imprisonment, even if they are considered safe and effective:

> P1 (Premise 1): Accept the *Claim of Moral Objectionability* based on M.
>
> P2 (Premise 2): Accept the *Claim of Criminal Liability*
>
> P3 (Premise 3): Accept the *Objection of Safety and Efficacy*
>
> P4 (Premise 4): Adopt the *Assumption of Safety and Efficacy*
>
> P5 (Premise 5): CNIs raise unique ethical issues because they involve direct intervention in the brain, a concern not applicable to TPs.
>
> Conclusion (C): Even with the hypothetical safety and efficacy of CNIs (P4) and acknowledging the moral implications of criminal acts (P2), CNIs' DBI brings up ethical issues that challenge their alignment with TPs.

In simpler terms, assuming CNIs are safe and effective, the way they directly change the brain still raises ethical concerns in itself. Unlike jail or other punishments that influence behaviour indirectly, CNIs work by directly altering brain functions. This direct approach brings

up ethical questions we do not face with traditional punishments, making CNIs ethically distinct and challenging their use, even if they are justified in other respects.

Response—The Parity Principle

The *Parity Principle* counters the DBI Objection by arguing that the ethical focus should be on the effects of neurointerventions, not their direct method of application. This principle suggests that if the outcomes of CNIs and traditional punishments are ethically comparable, the directness of CNIs should not be a defining ethical issue:

> P1 (Premise 1): Accept the *Claim of Moral Objectionability* based on M.
>
> P2 (Premise 2): Accept the *Claim of Criminal Liability*
>
> P3 (Premise 3): Accept the *Objection of Safety and Efficacy*
>
> P4 (Premise 4): Adopt the *Assumption of Safety and Efficacy*
>
> P5 (Premise 5): Acknowledge CNIs involve DBI.
>
> P6: (Premise 6): Argue, based on the Parity Principle, that CNIs' direct intervention does not necessarily make them more objectionable under metric M or distinct in a morally relevant way from TPs.
>
> Conclusion (C): Given the assumption of CNIs' safety and effectiveness (P4), along with the criminal liability incurred by offenders (P2), and the validity of the Parity Principle (P6), CNIs can be ethically justifiable and align with established theories of punishment P, challenging initial objections based on metric M.

In simpler terms, the argument suggests that what really matters ethically is not how CNIs work (directly on the brain) but their outcomes. Traditional punishments and CNIs might affect the brain in different ways—one indirectly and the other directly—but if the end results are similar, both direct and indirect interventions lead to similar outcomes—they influence the brain's operations, or "what neurons fire when and how." (Greely, 2008, p. 1134; Levy, 2007, p. 62; 2020, pp. 34-35). So CNIs should not be dismissed on the basis of their directness alone. The Parity Principle thereby challenges the notion that the directness of CNIs inherently brings unique ethical concerns, advocating for a focus on the effects rather than the means.

Ongoing Debate: Denying and Defending the Parity Principle

The debate on the Parity Principle (PP) focuses on the ethical use of CNIs versus TPs. Critics highlight CNIs' unique issue of bypassing rational thought, unlike traditional punishments, acting "directly on the mainsprings of action, on emotions or other dispositions" (Harris, 2012, p. 294). In contrast, supporters argue that both CNIs and traditional methods impact rational capacities, making the concern not unique to CNIs. Whether through TP or otherwise, all of us are manipulated on a daily basis "in ways to which we do not consent, using indirect interventions." (Levy, 2020, pp. 44-45). This debate highlights the PP's role in discussing CNIs' ethics, distinguishing between universal and situation-specific objections. We will further examine the PP's implications for criminal justice reform.

Recap

We have explored the debate on using CNIs versus traditional punishments, paving the way for a deeper look into PEAs. Here is a brief recap:

- **Moral Objectionability and Criminal Liability**: We have considered the ethics of using CNIs based on criminal actions, considering the fairness of brain interventions as a form of punishment.

  Do different ethical considerations arise for CNIs in relation to those who have committed crimes?

- **Safety and Efficacy**: Doubts about the safety and effectiveness of CNIs were addressed through scenarios that imagine CNIs as both safe and effective.

  Can we trust that brain interventions are safe and will work as intended? *If not, can we simply assume this is the case and theorize on this basis?*

- **Direct Brain Intervention Objection**: This objection raises questions about the ethicality of directly altering the brain (DBI) with CNIs compared to the indirect effects of traditional punishments:

  Does directly changing the brain with CNIs raise different ethical concerns than traditional punishments?

- **Debate on the Parity Principle**: Explored whether CNIs' direct impact on the brain is ethically distinct from the indirect impact of traditional punishments:

Are CNIs really ethically different because they directly affect the brain, unlike traditional punishments?

Punishment Equivalence Arguments

PEAs critically explore the use of CNIs in criminal justice, questioning if they can ethically match or even surpass TPs. At their core, PEAs probe whether CNIs and TPs could be considered morally equivalent or better than TPs under certain conditions.

Key Steps of PEAs:

- They begin with the *Motivating Question*: If we justify TPs, why not CNIs?
- They examine a potential moral objection (or objections) to CNIs, suggest a specific ethical metric (M) for evaluation, and incorporate concepts like criminal liability, presumed safety and efficacy of CNIs, and the Parity Principle.
- PEAs aim to demonstrate a moral equivalence between CNIs and TPs based on the claims of *Criminal Liability*, *Conditional Efficacy*, adoption of the *Parity Principle*, and a *Claim of Equivalence*.
- This leads to provisional implications that challenge the outright rejection of CNIs if TPs are accepted under similar ethical considerations.

In other words, PEAs encourage us to reconsider our automatic negative views of CNIs. If CNIs can deliver the same or better results than TPs without additional harm, dismissing them simply because they are new and unfamiliar is unjustified. This stance promotes a thoughtful discussion on integrating CNIs into criminal justice, guided by moral and ethical standards.

Formalizing PEAs[3]

To illuminate the complex ethical discussions around CNIs in criminal justice, this section organizes PEAs into a formal structure. By categorizing the discourse into four principal premises, each thoroughly examined in separate chapters, we achieve a clearer dissection of these multifaceted arguments. This approach simplifies the navigation through the nuanced moral debates surrounding CNIs, facilitating a more accessible and comprehensive exploration. Additionally, we survey notable examples from the literature to further contextualize these arguments, acknowledging the diversity of ways in which they can be structured and understood.

Preliminary: Metric and Moral Objection

PEAs hinge on the *Foundational Question*: If TPs are justified, why not consider CNIs? The discussion starts by identifying a specific moral standard (metric M) to assess the ethical standing of CNIs, leading to the initial *Claim of Moral Objectionability*.

Premise 1—The Claim of Punishment

> **Premise 1 (P1)**: If an individual commits a criminal act, this act incurs moral and, subsequently, criminal liability, which justifies the use of TPs based on established theories of punishment (P), *prima facie* justifying CNIs on similar grounds.

Premise 1 opens the discussion by recognizing the necessity of punishment for criminal acts, which may extend to the use of CNIs, given an initial acceptance of punishment's justifiability. However, it stops short of fully justifying CNIs, focusing instead on their potential equivalence with TPs. This approach introduces a foundational acceptance—that punishment, in some form, is warranted—while leaving room for deeper inquiries into whether CNIs and TPs are morally or practically equivalent.

While we might agree that punishment is necessary, we need to carefully examine how brain interventions compare to TPs like jail. Accepting that punishment is justified does not automatically mean our current ways of punishing people are right. This raises a key question: If we find brain interventions are as justifiable as traditional punishments, how does this compare with the reality of how we currently punish people?

The Conditional Efficacy Claim

> **Premise 2 (P2):** For the sake of argument, assume CNIs are both safe and effective in addressing criminal behaviour despite the lack of empirical evidence to this effect currently.

For this premise, we are asked to imagine a scenario where brain interventions are perfectly safe and effective, purely for the sake of argument. This helps us think about the ethical side of using CNIs if they work perfectly. However, it is important to remember that this is just a theoretical exercise to help us explore all of the potential ethical implications. This does not mean that CNIs are currently safe and effective in real life. The question this premise raises is

whether imagining a perfect scenario is helpful for our actual decision-making about using CNIs. Even if we accept this ideal scenario, we still need to consider what it really means to use CNIs in reality as part of our actual punishment practices.

The Parity Principle

> **Premise 3 (P3):** Accept, provisionally, the Parity Principle, positing that if CNIs are safe and effective, the ethical evaluation of CNIs and traditional punishments (TPs) should be judged primarily by their outcomes, such as deterrence and rehabilitation, rather than the methods used.

The Parity Principle posits that the ethical evaluation of CNIs and TPs should be based on their outcomes, such as deterrence and rehabilitation, rather than the methods used to achieve these outcomes. The principle suggests that if CNIs and TPs lead to similar results, they might be considered ethically equivalent, regardless of their differing methods.

Real-world Application: It evaluates how the principle functions within practical, real-world scenarios, acknowledging that outcomes are a critical measure of ethical equivalence.

Modal Analysis: It extends into a more theoretical realm, considering all possible scenarios to determine if the principle holds universally. This includes:

- *Necessity*: Testing whether the ethical equivalence of CNIs and TPs would apply in every conceivable situation, suggesting that similar outcomes always justify ethical equivalence.
- *Contingency*: Exploring specific conditions where the Parity Principle may not apply, such as if CNIs lead to unique, adverse effects not associated with TPs, indicating that equivalence might only be conditional.

It is argued that critics of the Parity Principle must demonstrate that any differences in the methods of CNIs and TPs, such as direct brain interventions by CNIs, have fundamental ethical implications that potentially challenge the universality of the principle. If the Parity Principle can be validated as a universal truth, it supports the notion that various correctional strategies could be ethically equivalent if they yield similar outcomes.

However, if intrinsic ethical differences between CNIs and TPs are identified—even assuming their safety and effectiveness—it suggests that the principle's application may be contingent and not universally valid.

The Parity Principle is central to PEAs, promoting a methodological stance that emphasizes outcomes in determining the ethical equivalence between CNIs and TPs. This dissertation critically evaluates the Parity Principle, scrutinizing both its methodological validity and its purported universal applicability. This examination transitions into a deeper inquiry into the Claim of Equivalence, challenging the sufficiency of outcome-focused assessments in fully addressing ethical considerations.

Claim of Equivalence

> **Conclusion (C)**: If CNIs are assumed safe and effective, and if we judge their impact based on outcomes (as per the Parity Principle), then their ethical standing in response to crime is at least equivalent to, if not better than, that of TPs, all measured against M.

This premise draws a direct line of comparison between CNIs and TPs, based on their potential outcomes. It challenges us to reevaluate CNIs in light of their ability to achieve the goals of punishment potentially more safely or effectively. While the Parity Principle and the Claim of Equivalence both prioritize outcomes, they operate differently within the context of PEAs. The former provides a broad evaluative framework, while the latter offers a more targeted comparison.

In simpler terms, The Parity Principle lays the groundwork, emphasizing outcomes over methods for CNIs and TPs, akin to creating a blueprint. The Claim of Equivalence then details this blueprint, directly comparing the two based on their effects in light of particular metrics M. However, in this dissertation, we challenge this foundation, especially the broad application of the Parity Principle and what implications this has for the Equivalence Claim. The question is whether CNIs truly align with TPs ethically, given their unique brain interactions, and whether theoretical comparisons hold amid real-world technological and ethical contexts.

Considerations—Reassessing Moral Objections and Considering Practical Viability

> **Consideration 1 (C1)—Provisional Reassessment:** Assuming CNIs are both safe and effective, we should reevaluate the initial moral objections to CNIs based on metric M. This reassessment does not outright endorse CNIs but opens the door for further consideration, suggesting that under ideal conditions, CNIs might not be as ethically problematic as initially thought.

The *Provisional Reassessment* suggests revisiting our initial moral objections to CNIs under the assumption that they are safe and effective. This conclusion does not outright endorse CNIs but opens the door to reevaluation, suggesting that under ideal conditions, CNIs might not be as ethically problematic as initially thought. It is a call to reconsider CNIs in a more favourable light, provided they meet certain ideal standards of safety and efficacy.

> **Consideration 2 (C2)—***Practical Viability Evaluation*: The examination of CNIs under ideal conditions brings forth considerations that might inform discussions on their permissibility and practical application within criminal justice systems.

This *Practical Viability Evaluation* moves the discussion from theoretical speculation to practical application possibilities, questioning how CNIs if assumed safe and effective, might be integrated into existing legal and societal frameworks based on their ethical evaluations. While the first consideration offers a provisional reassessment of CNIs, the second prompts further exploration of their potential practical integration and viability.

The journey from C1 to C2 draws on earlier premises about the justification for TP, the hypothetical safety and efficacy of CNIs, and ethical comparisons made through the Parity Principle. Together, they propose a transition from viewing CNIs in a new ethical light to pondering their practical ethical viability, emphasizing the nuanced journey from ethical theory to actionable insights in criminal justice reform.

Literature Review

Before proceeding, we will consider some of the contemporary literature on PEAs. We start by exploring a seminal 2014 article by Thomas Douglas that has been a cornerstone in the

ethical discussion of neurointerventions in criminal justice. Following this, we delve into more recent works that build on and diverge from Douglas's ideas, offering a spectrum of views that enrich our understanding of CNIs' ethical landscape. This approach allows us to see how the debate has developed, from foundational insights to the latest thinking.

Dissecting the Foundations: An Examination of Douglas's Argument on Moral Liability and CNIs

Douglas's 2014 article is a key contribution to the PEA regarding CNIs in criminal justice. He examines consent, moral liability, conditional efficacy, and the parity principle, leading to a provisional conclusion that invites reconsideration of CNIs' ethical justification. This analysis provides a structured framework for evaluating the complex ethical implications of CNIs, serving as a foundation for further exploration within the PEA discourse.

Why Not? – Moral Objectionability and the Issue of Consent

Douglas (2014) addresses the foundational question of PEAs: if TPs are justifiable, why should CNIs be excluded out of hand? He identifies a *Claim of Moral Objectionability* based on the contentious issue of consent, challenging the dominant assumption that CNIs are inherently objectionable due to potential consent violations (p. 104). Through the "Consent Requirement," Douglas questions whether medical interventions, like CNIs, necessitate an offender's valid consent to be deemed ethically permissible. This examination serves as a prototypical exploration within the PEA framework, embodying its objective to critically reassess and potentially refute initial moral objections against the use of CNIs, specifically through reevaluating the consent criterion as a key metric (M) of moral objectionability.

Punishment

Douglas tackles the idea of "moral liability" in his Punishment Claim, suggesting that crimes justify state-imposed punishments without needing the offender's consent, as stated, "the state may permissibly do things to criminal offenders without their consent... it could not do to others" (p. 105). This moves from the idea of moral to criminal liability, justifying punitive measures like jail. His argument, which sees criminal actions as meriting specific state responses, including potentially neurointerventions, connects closely with the broader discussion of Punishment Equivalence Arguments (PEAs). Douglas uses this claim to discuss the rehabilitative

aims of punishment, P, integrating the concept of moral liability with the justification for using such measures.

Conditional Efficacy

Douglas navigates the Safety and Efficacy Objection with a Conditional Efficacy Claim, proposing a thought experiment about a safe and effective neurointervention described as a "medical corrective" (2014, p. 104). He argues that if CNIs can be assumed to be safe and effective, then they hold potential for significant rehabilitation benefits. This assumption is not a claim of current viability but a conditional setup to explore CNIs' ethical landscape. It reflects a key PEA discourse, examining the conditional ethical acceptability of CNIs based on their potential for safe and effective rehabilitation. This hypothetical approach highlights Douglas's contribution to PEA discussions, focusing on the broader implications and ethical considerations of using CNIs in criminal rehabilitation under ideal conditions.

The Parity Principle

Douglas challenges the idea that CNIs uniquely bypass rational thought, comparing their impact on the mind to that of TPs like incarceration. He argues that both methods can significantly alter an individual's mental state, questioning the unique ethical objections to CNIs. By invoking the Parity Principle, Douglas suggests that the ethical evaluation of CNIs should focus on their outcomes and rehabilitation effects rather than the method of intervention. This approach calls for a balanced, ethical assessment of CNIs, emphasizing their potential benefits and comparable impact to conventional punitive measures (see also Douglas, 2018).

The Equivalence Claim

Douglas concludes his examination with the Equivalence Claim, bringing together arguments about moral liability, the assumed safety and efficacy of CNIs, and the Parity Principle. He introduces the "Robustness Claim" to compare the ethical weight of different rights infringed by CNIs and traditional punishments. This analysis focuses on whether objections based on consent and bodily integrity hold more ethical weight than those related to traditional punishments like incarceration.

Douglas suggests that when considering the moral accountability of offenders, the hypothetical safety and efficacy of CNIs, and their effects on rational capacities, the ethical

distinction between CNIs and traditional punishments diminishes. He argues that CNIs, under these scrutinized conditions, might not be more ethically problematic than traditional punitive methods, challenging the assumption that CNIs are inherently less permissible due to their direct brain intervention.

After exploring Douglas's nuanced approach to CNIs, it is clear his argument reflects the intricate relationship between the Parity Principle and the Equivalence Claim within the framework of PEAs. While Douglas employs the Parity Principle to establish a broad, outcome-focused basis for evaluating CNIs, his progression to the Equivalence Claim reveals a deeper, targeted analysis that scrutinizes the ethical weight of CNIs against traditional punishments—reflecting the PEA formalization above.

Considerations—Lowering Moral Barriers

Douglas's concluding analysis suggests a conditional openness to the ethical use of CNIs. He articulates this by challenging the Consent Requirement, proposing that CNIs "could in principle be justified" and should not be automatically ruled out due to their nonconsensual nature (p. 120). This stance reflects the core of PEAs: to question and potentially diffuse moral objections to CNIs without necessarily asserting their outright ethical permissibility.

However, Douglas carefully qualifies this position, noting "it may not follow that the compulsory imposition of medical correctives is in fact justified" (p. 107), indicating a provisional rather than a conclusive endorsement of CNIs. His further mention of mental integrity rights showcases the nuanced debate within PEAs, emphasizing the complexity of fully justifying CNIs.

In essence, Douglas does not advocate for the immediate use of CNIs but encourages a reconsideration of our ethical reservations about them. His approach exemplifies the PEA methodology by inviting continued exploration of CNIs' ethical considerations, positioned within a broader discourse on criminal justice reform.

PEAs—The Ongoing Debate

Since Thomas Douglas's seminal work in 2014, the discourse surrounding CNIs within criminal justice systems has significantly evolved, transcending initial concerns of consent to address a wider range of ethical considerations. This shift has heralded an era of enriched

dialogue, with my dissertation navigating these expanded debates to dissect the complex ethical fabric of CNIs.

The last decade has witnessed an enriching proliferation of perspectives on CNIs, each bringing to light new facets of their ethical dimensions. This development is encapsulated in the 2018 "Treatment for Crime" series. (David Birks & Thomas Douglas, 2018), and a further collection of articles framed as PEAs aimed at "Diffusing Objections" against CNIs.

Key scholarly contributions in the form of PEAs have dynamically shaped the conversation around CNIs, broadening the ethical lens through which we view these interventions. From Jeff McMahan's (2018) exploration of CNIs grounded in moral liability and harms to Kasper Lippert-Rasmussen's (2018) critique of the self-ownership thesis via the extended mind thesis, Birks and Buyx's (2018) of punitive intentions and Emma Bullock's (2018) analysis of moral paternalism, alongside Stefano Fuselli's refinements, with emphasis on mental integrity and cognitive liberty (2021). The debate continues.

At the core of these diverse discussions lies a shared structure reflective of PEAs: the foundational "why not" question, the identification of specific moral objections (M) — ranging from harms and self-ownership to moral paternalism and mental integrity — and the engagement with the principles of punishment justification (P), conditional efficacy, the Parity Principle, and the pivotal notion of equivalence. The majority of the arguments endorse provisional considerations, identifying the challenges that would need to be addressed before real-world implementation. We consider these arguments and lines of analysis throughout this dissertation.

Mapping the Terrain: Understanding Core Claims and Key Distinctions

Through a detailed look at PEAs, we have seen what they assert and their position in the broader debate. Before diving into the main chapters of this dissertation, it's important to outline this study's focus and introduce some critical distinctions to guide our discussion.

Scope and Core Arguments of the Dissertation—Negative and Positive Claims

In our detailed look at PEAs, we have seen what they assert and their position in the broader debate. Before diving into the main chapters of this dissertation, it is important to outline this study's focus and introduce some critical distinctions to guide our discussion.

Negative Claim: My central argument is that PEAs face numerous challenges that weaken their premises and question their overall validity—or preemptively challenge or block the provisional C1 and the practical considerations C2, questioning the guidance PEAs can provide for practical application. These challenges arise from PEAs' inability to align CNIs to TPs across various ethical and normative dimensions, notably in terms of 'mental freedom.'

Positive Claim: Beyond identifying flaws, this work asserts a possible ethical difference between CNIs and TPs, centred on the unique impact of CNIs on mental freedom. This difference suggests a fundamental departure from conventional punitive approaches and questions the equivalence claimed by PEAs, particularly because they do not fully address the concerns over the intrusion of neurointerventions into a domain of personal liberty traditionally safeguarded from state intervention.

In simpler terms, the *negative claim* highlights significant issues with PEAs, challenging their logical soundness and the proposed equivalency between CNIs and TPs. Akin to finding mistakes in a map that misguides us about the real layout of the land.

The *positive claim* goes further, pointing out the ethical distinction between CNIs and TPs, especially regarding mental freedom, not unlike realizing not only that the map is wrong but also that it misses crucial details necessary for a complete and safe journey through the terrain. In taking aim at each premise of PEAs, each chapter aims to support both the positive and negative claims.

Preliminary Distinctions

Transitioning from our discussion on PEAs and the negative and positive claims within the context of PEAs, it becomes imperative to delve deeper into the intricate layers of this debate. To navigate the complexities and ethical considerations of CNIs, we must first establish a few foundational distinctions. These will guide us through our analysis in the chapters to follow.

The Neuroscience of Ethics and the Ethics of Neuroscience

First is a distinction in the field of neuroethics, which is where this dissertation is situated.

The emergence of neuroethics can be traced back to roots in "neurophilosophy." (Churchland, 1989) through more refined discussions about *Bioethics and the Brain* (Glannon, 2007), culminating in its recognition as a distinct discipline near the turn of the century. This development led to the recognition of two distinct branches within neuroethics: the ethics of neuroscience and the neuroscience of ethics (Roskies, 2002; Roskies, 2020b).

The ethics of neuroscience examines ethical issues surrounding the development and use of neurotechnologies, including interventions in the brain, such as ethical standards and considerations that should govern the use of neurointerventions in the criminal justice system.

The *neuroscience of ethics* explores the neurobiological mechanisms underlying rationality, moral decision-making, and human nature. This branch aims to consider how advancements in neuroscience might cast light on the features that make us uniquely human. (Roskies, 2002, pp. 21-22) And "the kind of creatures we are" (Levy, 2007, p. 8), and "the social structures that we inhabit and create" (Roskies, 2020b).

Neuroethics weaves together two deeply interconnected strands: the ethics of neuroscience, which scrutinizes how we should ethically approach brain interventions, and the neuroscience of ethics, revealing how our brains process moral and ethical decisions. This blend forms the backdrop against which this dissertation unfolds, offering a nuanced lens to examine the ethics of neurointerventions in criminal justice. It is important, as I rely not only on the neuroscience of ethics in addressing issues—which is the predominant focus of PEAs—but expressly draw on the neuroscience of ethics in advocating more comprehensive assessments of PEAs.

Ideal and Non-Ideal Theory

Neuroethics, part of the wider field of applied ethics, requires us to distinguish between 'ideal' and 'non-ideal theory.' This distinction is crucial for our discussions, tracing themes throughout this dissertation.

Rooted in key texts like Rawls's *A Theory of Justice*, 'ideal theory' helps us imagine a just society, whereas 'non-ideal theory' tackles the hurdles of making such a society real amid existing injustices and inequalities (Rawls, 1971, p. 245; A. J. Simmons, 2010).

In neuroethics, the distinction between 'ideal' and 'non-ideal' theories borrows from this distinction. Ideal theory explores ethics in hypothetical, optimal settings, crafting frameworks from abstract principles while sidelining a spectrum of real-world variables. Conversely, non-ideal theory grapples directly with real-world ethical challenges, considering technological limitations, empirical data, societal dynamics, and the practicalities of ethical application. (Nadelhoffer et al., 2020, p. 196; Ryberg, 2020, p. 188).[4]

In discussing PEAs, we will see a tendency to use idealized scenarios where punishments or neurointerventions are assumed to be safe or at least minimally harmful and setting aside broader social or political implications. (Douglas, 2014c, p. 112; Greely, 2008, p. 1134; Levy, 2007, p. 72; Ryberg, 2020 Ch 3)—building on the Assumption of Safety and Efficacy and Conditional Efficacy.

This approach isn't wrong *per se*; it's about exploring possibilities—imagining what could be done to address urgent ethical issues or anticipate future ethical dilemmas in the fast-paced field of neuroscience. This way of thinking helps clear away some of our initial reservations, highlighting pressing concerns we should focus on as technology advances (Illes & Racine, 2005, p. 12; Stefano, 2021, p. 211; Farah, 2011a, p. 776; Vincent, 2014, p. 21; Ryberg, 2020, pp. 11-12).

However, there are inherent limitations to this approach. Some argue it "sets aside a wide range of important practical issues that are morally relevant for the adoption of laws" and "may obscure some of what matters ethically" (McTernan, 2018a, p. 284; Vallentyne, 2018b, p. 138): "Idealisation can only get us so far" (Birks & Buyx, 2018, p. 135). There is a risk ideal theorizing obscures potentially devastating real-world side effects that biomedical interventions may bestow upon individuals" (Focquaert et al., 2020, p. 142).

In short, the distinction between ideal and non-ideal theory is like the difference between planning in a perfect world and dealing with the messy realities of life. Ideal theory lets us dream up the best scenarios without worrying about obstacles, while non-ideal theory forces us to navigate the complexities and challenges we actually face.

Finally, I do not see PEAs as reflecting a binary approach. Their situated spectrum allows us to appreciate the diversity of assumptions and considerations that play into our ethical judgments. It's not just a matter of black and white but a gradient of grey, where each point on

the spectrum represents a balance between the theoretical ideal and the practical realities we encounter.

This nuanced approach is particularly valuable when analyzing PEAs—which often idealize CNIs—as it enables us to evaluate them from multiple angles, challenging not only the premises but the conclusions these arguments actually support. Ideal theory gives us the ethical ideals to strive for, but it is the non-ideal theory—with its focus on safety, efficacy, and real-world challenges—that grounds our analysis in reality, making the discussion around the ethical use of CNIs in criminal justice both comprehensive and relevant.

As I hope to show, if the aim of PEAs is to provide *Provisional Considerations* (C1), then concerns about idealization are assuaged to some extent. However, to the extent they endorse *Practical Considerations* (C2) and are to offer any actionable guidance to actual practice, the risk of idealization is greater. As such, throughout this dissertation, I identify issues with Idealization, further to both my negative and positive claims.

Moral and Normative Asymmetry

Finally, PEAs explore whether CNIs and TPs can be considered morally equivalent based on a specific moral framework, or 'metric M.' Douglas, for instance, delves into the concepts of consent and rights to argue for this equivalence. However, searching for equivalence also implies the notion of 'asymmetry'—relevant differences that must be addressed—highlighting that, despite efforts to demonstrate their ethical parity, significant differences between CNIs and TPs emerge, especially when broader ethical and practical considerations come into play.

Simply put, PEAs scrutinize if CNIs and TPs are ethically similar or not based on a chosen metric, 'M,' that outlines moral objections. This process is not just a comparison but a deep analysis of moral foundations and their practical implications.

To distinguish the sorts of considerations M and PEAs are interested in, this dissertation differentiates between what I describe as moral asymmetry and normative asymmetry. I take *moral* judgments to be based on broader first-order considerations of fairness, rights, and duties—based on what I will describe as *moral* reasons. On the other hand, *normative* judgments consider not only moral norms but also other types of norms, such as social norms and legal norms, that include not only *moral* reasons but prudential and *practical* reasons. (Brink, 2010; Dancy, 2023; Wedgwood, 2007, p. Ch 1).[5]

The concepts of moral and normative asymmetry bridge discussions between ideal and non-ideal theory, marking a transition from provisional considerations—challenging our intuitions— to practical considerations—the real-world impacts of CNIs.

For example, we might establish a moral equivalence between CNIs and TPs under idealized scenarios, arguing their ethical objections are similarly valid. However, normative asymmetry could persist due to real-world issues like safety, effectiveness, and how these interventions fit within existing penal systems. Even if PEAs are successfully calibrated to theoretically resolve moral asymmetry by sidelining these practical concerns—a separate issue I contest—they might not tackle the resulting normative asymmetry.

Consequently, even sound arguments that lessen moral objections might neglect broader practical and ethical considerations, underscoring the importance of examining both moral principles and pragmatic aspects in CNIs' ethical evaluation. This gap highlights the challenge of moving from hypothetical scenarios to actionable guidance in real-world contexts, ultimately questioning the comprehensive ethical justification of CNIs.

In other words, this distinction is important throughout the assessment of PEAs. Because equivalence is a central feature of PEAs, the distinction between moral and normative asymmetry allows us to carefully classify the sort of similarities or differences we are interested in, allowing us to maintain a focus on whether and to what extent PEAs are successful in establishing equivalence.

Recap

Before advancing to the chapter overviews, a concise recap is due to ensure clarity on the discussion thus far:

- **Introduction and General Debate**: We initiated our exploration with an inquiry into CNIs within the criminal justice context, probing the foundational moral objections and the "why not" question, leading to a reevaluation of our moral intuitions about CNIs based on liability, assumptions about safety and efficacy, and outcomes, not methods.

- **Structuring PEAs**: We dissected PEAs, delineating their core premises around punishment, CNIs' conditional efficacy, the Parity Principle, and their purported ethical equivalence with traditional punishments. This aimed to bridge moral objections to CNIs by positing their ethical comparability to traditional punishments under certain premises and considering purported provisional considerations about moral barriers and possible practical considerations about actionable real-world guidance.

- **Positive and Negative Claims**: This dissertation distinguishes between the 'negative claim,' critiquing PEAs for not convincingly demonstrating ethical equivalence, and the 'positive claim,' advocating for the ethical distinction between CNIs and traditional punishments, particularly regarding 'mental freedom.'

- **Further Distinctions:**

    - Explored the ethics of neuroscience versus the neuroscience of ethics, emphasizing the ethical exploration of neurotechnologies and the neurobiological basis of moral decisions.

    - Highlighted the divide between ideal and non-ideal theory, noting PEAs' tendency towards idealization, which may neglect practical ethical issues.

    - Introduced moral and normative asymmetry, questioning PEAs' ability to establish normative equivalence alongside moral equivalence, given wider ethical and practical considerations.

We situate discussions in the broader debate, consider the premises of PEAs successively, and circle back to these distinctions throughout the following chapters, and they will serve as a roadmap to carefully chart the court and ensure analytic clarity.

Summary of Chapters to Follow

Before we jump into Chapter 1, let us provide a short summary of the four chapters that follow, what they attempt to establish, and how they aim to advance both the negative and positive claims.

Punishment

Punishment—Chapter 1 of my dissertation, focusing on the Punishment Claim (P1), sets the stage for a nuanced discussion on the ethical implications of CNIs within the framework of PEAs. This chapter lays a foundational premise that criminal behaviour renders individuals liable to punishment, justified by specific theories of punishment (P), and explores the theoretical possibility of CNIs serving as a form of such punishment.

Accept *Prima Facie* Acceptability of Punishment and Penal Justification—I begin by acknowledging the necessity of punishment for serious offences. This acceptance sidesteps the broader debate on the validity of various theories of punishment to focus on the ethical consideration of CNIs within these universally condemnable acts. By suggesting that CNIs could, in theory, align with traditional theories of punishment in an idealized scenario, I set the

groundwork for further examination while clarifying, as I argue in later chapters, that this acceptance does not support the Equivalence Claim and equate CNIs and traditional punishments morally.[6] In other words, let us accept that we are justified in punishing in at least some cases and open the discussion by accepting that CNIs could, at least theoretically, be justified as punishment.

Challenge Reliance on Idealization and Baseline Objection—The chapter then delves into the critique of idealizing punishment and its practical implications, emphasizing the disconnect between idealized conceptualizations and the real-world effectiveness of penal practices. This critique focuses on the difference between ideal and non-ideal theory and moral and normative asymmetry, highlighting the practical and ethical challenges inherent in aligning theoretical justifications of punishment with the realities of the current penal system, thus questioning the viability of neurointerventions as a straightforward solution to criminal behaviour.

Essentially, the chapter questions the leap from theoretically endorsing punishment, and by extension neurointerventions, to their practical and ethical viability in today's penal landscape. It suggests that while the idea of using advanced techniques to address criminal behaviour is compelling in a perfectly just world, the realities of our flawed penal system introduce serious challenges that undermine the straightforward application of such idealized solutions.

Problems with Isolating The Criminal Brain and Treating Crime— In addressing the complexities of criminal behaviour and the potential misguidance of focusing solely on CNIs for 'treating' crime, the chapter then underscores the importance of considering broader societal and evolutionary influences on punishment practices. This discussion builds on the neuroscience of ethics and advocates for a more holistic approach to criminal justice, one that transcends the mere application of new technologies to address deeper, systemic issues.

Relevance for Negative and Positive Claims—By critically evaluating the justification for punishment against the backdrop of flawed penal systems and the speculative nature of CNIs' efficacy, Chapter 1 aims to both challenge the foundational assumptions of PEAs (negative claim) and promote a more thoughtful, ethical consideration of CNIs within the justice system (positive claim). This approach advocates for a reevaluation of our justice system's reliance on punitive measures, suggesting a shift towards more humane and effective solutions to crime.

Conditional Efficacy

Conditional Efficacy—Chapter 2 considers the 'Conditional Efficacy' claim within PEAs, positing that CNIs are presumed to be safe and effective for the sake of argument (P2). This approach abstracts from the complexities of current neurotechnological capabilities and uncertainties, placing the discourse within an idealized framework that assumes a hypothetical scenario of optimal safety and efficacy.

The question addressed in this chapter is whether it is suitable to assume CNIs work well without really looking into the current technology's limitations. Is this an acceptable approach based on the present science? How far from reality are these assumptions? What implications does this have for what we can learn from PEAs or the guidance they can offer in the real world?

Idealization is Acceptable But Demands Significant Caution—Acknowledging the role of idealization in neuroethics; the chapter recognizes the search for 'in-principle' distinctions between neurointerventions and traditional methods of influencing behaviour as a common and not inherently flawed approach. (Douglas, 2014c, p. 112; Greely, 2008, p. 1134; Levy, 2007, p. 72; Ryberg, 2020 Ch 3). However, it cautions against the potential oversimplification of complex issues and the dangers of prematurely implementing CNI regimes based on overly optimistic theoretical assumptions.

Explore Brain, Mind, and Current Limitations: By examining the current state of neuroscience, including the biological underpinnings of morality and criminal behaviour and the capabilities of existing CNI technologies, the chapter aims to ground its analysis in the real-world context of the neuroethical discussion. This exploration serves as a foundation for subsequent discussions on human rationality and the concept of mental freedom, integrating insights from the neuroscience of ethics to inform the debate on the ethical implications of CNIs.

Critique of Ideal Theorizing Given Shortcomings of Current Science—This chapter concludes by scrutinizing the claim of Conditional Efficacy against the backdrop of current neuroscientific evidence, suggesting that such assumptions push the discourse into the realm of speculative science fiction rather than practical ethical consideration.

Support for Positive and Negative Claims—This critical perspective underscores the chapter's support for the dissertation's *negative claim* by highlighting the significant ethical and practical challenges in equating CNIs with traditional punishments without addressing the substantive issues within our justice system. Furthermore, the chapter contributes to the *positive*

claim by emphasizing the need for careful and principled consideration of neurointerventions. By pointing out the normative asymmetry and advocating for a more nuanced understanding of the ethical landscape surrounding CNIs, the chapter calls for a thoughtful approach that transcends mere technological innovation to address the underlying societal and systemic challenges in criminal justice.

In summary, Chapter 2 interrogates the Conditional Efficacy claim of PEAs, revealing the profound gap between theoretical assumptions of CNIs' safety and efficacy and the current realities of neurotechnology. Through a detailed examination of the scientific and ethical dimensions of neurointerventions, the chapter both challenges the adequacy of PEAs and advocates for a more responsible and ethically informed exploration of CNIs within the context of criminal justice reform.

The Parity Principle

The Parity Principle: Ends not Means—The Parity Principle chapter delves into the heart of PEAs by scrutinizing the belief that the direct brain intervention of CNIs is equivalent to traditional punitive methods when both ultimately influence brain function. The principle asserts that if CNIs, presumed safe and effective, achieve the same ends as TPs, the means of achieving these ends (CNIs) should not carry ethical weight. (Levy, 2007, 2020). This chapter challenges this assumption by exploring the complex debate around the Parity Principle, focusing on the criticism that CNIs bypass rational capacities, potentially subverting freedom in a way TPs do not.

Surveying the Debate on the Parity Principle: Direct and Indirect—In the first part of the chapter, I survey the deeply divided debate surrounding the parity principle in the literature. Critics argue direct interventions are distinct because, unlike conventional forms of punishment, they have the unique ability to 'circumvent rational capacities' and induce changes without the individual's conscious awareness. In this sense, they are sometimes described as 'freedom subversive.' (Bublitz, 2020a, 2020b; Bublitz & Merkel, 2014; Focquaert & Schermer, 2015; Harris, 2011, 2012, 2013a, 2013b, 2014a; Shaw, 2012). I also consider responses.

Not Two, but Three Ways of Changing the Mind—Next, by leveraging advancements in neuroscience, the chapter argues against the binary view of mind change (direct versus indirect), introducing a third category: internal mentation, where individuals can independently

modify their mental states. This addition highlights overlooked aspects of human rationality, such as episodic memory and embedded cognition, suggesting that the Parity Principle fails to capture the entirety of how our minds work.

Implications for the Parity Principle and PEAs—While the Parity Principle serves as a foundational methodological aspect of PEAs, emphasizing outcome over method, this chapter prepares to challenge its broad applicability. Specifically, we will explore concerns that CNIs, through their direct intervention on the brain, might present unique ethical issues not accounted for by a simple comparison of outcomes with traditional punishments. This critique sets the stage for a more detailed examination in the following chapters, especially when considering the nuanced implications of the Equivalence Claim and the real-world application of CNIs. Our exploration suggests that the ethical evaluation of CNIs requires a more intricate analysis than what the Parity Principle alone can provide, leading us into a deeper debate on the nature of punishment, autonomy, and the potential for self-directed change or 'internal mentation.'

Support for Positive and Negative Claims—In summary, this examination of the Parity Principle within PEAs highlights its critical role in arguing for the ethical equivalence of CNIs and TPs. By identifying overlooked dimensions of rationality and challenging the principle's assumptions, the chapter supports the dissertation's negative claim by pointing out a major analytic oversight in PEAs, which both challenges the Parity Principle premise and has further implications for the equivalence claim. Moreover, it bolsters the positive claim by emphasizing the need for a thorough understanding of rationality before hastily adopting CNIs, thereby advocating for a more cautious approach to their implementation in criminal justice.

In simple terms, if the parity principle is wrong, that there are only two ways of changing the brain and mind, then it is difficult to defend the idea that CNIs and TPs do not raise different concerns. This means PEAs are not logically sound and also that CNIs might pose different threats that make it difficult to explain or defend how their effects are the same. This sets the stage for a deeper evaluation of the implications of the Equivalence Claim in the subsequent chapter, focusing on the significance of mental freedom and autonomy.

The Claim of Equivalence

The Equivalence Claim—The final chapter dives deep into the 'Claim of Equivalence,' a core part of PEAs, comparing the effects of traditional punishments and CNIs—attempting to identify equivalence, or the 'lesser of two evils' (Bublitz, 2018).

Identify Relevance of Selecting a Moral or Ethical Metric—I begin the Chapter by reiterating the importance of PEAs clearly defining and choosing a specific moral or ethical standard (metric M) to evaluate the arguments for and against neurointerventions compared to traditional punishments. This clarity is essential for understanding the arguments' strengths and weaknesses and for addressing significant concerns, especially those related to 'human freedom.' This challenge spans various fields of study and theories, emphasizing the need to consider both moral principles and practical concerns in evaluating these arguments.

This is crucial because the given metric, M, has implications for the conclusions PEAs are able to support, including assessing theoretical objections and whether 'lower moral barriers' (C1) of the stronger claim they lend support for practical applications in real-world scenarios (C2). I argue a thorough examination of 'human freedom' as a metric is deemed essential before any significant reforms in penal practices are considered, given the diverse interpretations and implications across various academic and practical fields.

In simple terms, I start by emphasizing how clear we need to be when using PEAs to compare brain interventions with standard punishments, especially about what we are measuring. It is complicated because this conversation touches on many different areas, from moral theory to political philosophy, applied ethics and law, so we need to think carefully about all the reasons why one might be better or worse than the other and how much terrain we must explore before we are confidently moving forward in the real world. This is important for making sure we do not rush into changing how we handle punishment without understanding all the consequences.

Introducing Thought Experiments—I then consider two thought experiments from the neuroethics field to shed light on the nuanced debate about freedom and its implications for punishment and crime. John Harris's 'Freedom to Fall' (Aristotle, 2009; Harris, 2011, p. 104; 2013b; 2014b, p. 75; John Harris, 2016) and Persson and Savulescu's 'God Machine' (Savulescu & Persson, 2012) are used to question the ethics of controlling behaviour, thus challenging simple comparisons between neurointerventions and traditional punishments.[7] These scenarios prompt a deeper reflection on human rationality and freedom, suggesting that the issues

at play in punishment equivalence arguments are complex and resistant to straightforward equivalences.

Put another way; the chapter then uses two imaginary scenarios from science and philosophy to show just how complicated it can be to think about using new brain technologies for punishment. These examples help us see why it is not so simple to say brain interventions are the same as or better than old ways of dealing with crime. It is all about understanding what it really means to be free, how changes to our brains through CNIs might affect that freedom, and whether the way they do so sets them apart from how we normally punish.

Human Freedom: Domains of Inquiry—Next, the chapter delves into the complex issue of human freedom by examining it through various scholarly lenses: first-order moral theory, political philosophy, public policy, bioethics, and law. This multi-faceted approach is taken to thoroughly investigate whether CNIs present ethical concerns distinct from traditional punishments TPs, particularly accounting for 'mental freedom,' and how it might fit into various theories and metrics across these domains. I consider how mental freedom might be relevant to explaining how CNIs might be seen as wrong at different levels on different theories. The exploration acknowledges that, despite extensive theoretical and empirical grounding in human rationality, it remains uncertain whether we can fully address and dismiss concerns about the unique threats CNIs pose to human freedom, potentially creating a significant ethical gap between CNIs and TPs. The challenges identified within ideal theorizing about the Parity Principle indicate that resolving these concerns necessitates further empirical and theoretical work, especially considering the intricacies encountered in applied and non-ideal theoretical contexts. These domains reveal persistent issues of normative asymmetry—differences grounded in practical or prudential considerations—that are unlikely to be resolved in the near future.

In other words, we look at the big question of freedom from many angles, like ethics, politics, and law, to see if brain-changing punishments are really the same as or different from other types of punishment. It points out that we can not yet say for sure that these new methods are okay because they might threaten important aspects of being human, like our freedom and dignity, in ways we do not fully understand yet. So, we are encouraged to think more about what "freedom" really means in these debates, especially considering how advancements in brain science could change our views on punishment and freedom.

Challenges to PEAs: Positive and Negative Claims—The conclusion for the equivalence chapter effectively synthesizes the dissertation's examination of PEAs, emphasizing the nuanced challenges these arguments face when scrutinizing the ethical and practical dimensions of CNIs compared to TPs. It underscores that PEAs often oversimplify the complexities inherent in applying neurointerventions within the justice system, particularly by not adequately addressing the concept of 'mental freedom.' The analysis reveals that while PEAs aim to establish ethical equivalence between CNIs and TPs, they falter by not considering the full spectrum of implications such interventions have on human autonomy and rationality. This oversight is significant, as 'mental freedom' emerges as a crucial metric for evaluating the moral and ethical standing of CNIs.

Despite potential benefits, the concept's broad applicability and interpretive flexibility pose challenges. Grounded in a scientific understanding of human rationality, 'mental freedom' underscores profound asymmetries between CNIs and TPs, challenging the foundational claims of PEAs. This conclusion affirms the dissertation's negative claim regarding the logical flaws within PEAs and advocates for a cautious, principled exploration of neurointerventions. It calls for further inquiry into 'mental freedom' and its broader ethical implications, proposing a forward-looking agenda for neuroethics research that carefully navigates the intersection of technological innovation, human dignity, and the preservation of autonomy.

Put another way, we wrap up the discussion on whether brain interventions and traditional punishments are basically the same for moral or ethical purposes. I argue that these arguments do not cover everything or close off every avenue when it comes to our freedom to think and make choices. While brain interventions might help in some ways, we need to think hard about their impact on what makes us who we are. The big takeaway is that we need to tread carefully with new technologies that could affect our freedom and that there's still a lot to figure out about how to do this ethically. In fact, when we look closer, there are many reasons to believe CNIs do pose distinct threats to freedom that TPs do not. It is a call to keep digging into these tough questions as we move forward.

Concluding Remarks—Mental Rights and a Path Forward

In the conclusion chapter, I argue this dissertation by claiming to have identified considerable challenges for PEAs—in support of my negative claim. However, I also

acknowledge that I have not decisively grounded a positive claim that would generalize across all domains or potential applications. Analysis is by no means exhaustive. The complexity of these issues necessitates more in-depth examination, particularly concerning the concept of 'mental freedom' and its relevance to various contexts, such as rational agency, authenticity, and personal identity.

A comprehensive assessment of real-world implications calls for a nuanced understanding of 'freedom,' balancing penal theoretical considerations with specific neurointervention effects. It also requires recognition of the potential for certain neurointerventions to enhance freedom on some conceptions by restoring capacities necessary for autonomous agency. While this dissertation hints at potential avenues for future inquiry—notably, the discourse on moral or legal rights over the mind—it is clear that the journey is far from over. The path that lies ahead calls for an exploration of issues beyond the scope of this research.

The overall aim of this dissertation is to begin an important conversation about the ethical implications of using brain interventions for punishment. It delves into leading arguments but acknowledges there's much more to understand, especially regarding how these interventions might affect our fundamental freedoms. The possibility that some interventions could enhance freedom by addressing brain issues is intriguing, but it's clear we're just starting to grasp the broader implications. It's vital to continue these discussions carefully to avoid overlooking critical aspects of mental freedom and rights. This journey of inquiry is just beginning, emphasizing exploration over immediate conclusions.

In the end, I suggest that, at minimum, the dissertation serves to illuminate the myriad issues surrounding punishment equivalence and novel neurointerventions and to identify areas ripe for further exploration—and those that warrant the most careful consideration. And all tie, in many important ways, to notions of 'mental freedom'—a concept as ubiquitous as it is compelling. But the ultimate destination lies further ahead, promising richer insights and deeper understanding. As Stevenson (1881) remarked in his *Virginibus Puerisque*, 'for to travel hopefully is a better thing than to arrive, and the true success is to labour.' And so we begin by turning to the realm of criminal punishment.

1

The Punishment Claim

An Irrational System of Punishment

Introduction

In 2012, Adam Capay, a 19-year-old Indigenous Canadian, found himself facing murder charges for the fatal stabbing of a fellow inmate. His life trajectory was shaped by intergenerational traumas. Growing up in abject poverty and a fractured home, Capay, like his father before him, endured significant neglect and experienced instances of physical and sexual abuse. The most chilling of these memories dates back to when he was just ten years old, when his own father held a loaded gun to his head, demanding Adam pull the trigger.

Capay turned to alcohol, drugs, and inhalants at a young age, seeking solace from his troubled reality. Unfortunately, his path soon intersected with the justice system, and at the time of his charges, he grappled with a range of psychiatric disorders, including depression, self-harm, and suicidal ideations. Despite awareness of these issues by prison authorities, Capay was subjected to an extended period of solitary confinement while awaiting trial, lasting an alarming 1647 days. This duration exceeded the United Nations' prescribed limit for such treatment by over 100 times, which is considered tantamount to torture.[8] During his confinement, Capay's health deteriorated significantly, leading to hallucinations and numerous severe acts of self-harm. He inserted a pencil through his right cheek, slashed his arm with a razor blade, and repeatedly banged his head against the cell wall until he bled. Disturbingly, near the end of his confinement, it was reported that Capay had begun to lose his ability to speak.

In a remarkable decision, a Justice from the Ontario Superior Court in Canada 'stayed'—or dismissed—the murder charges, finding no other remedy could adequately address the harm caused to the integrity of the justice system (Gilmore, 2016).[9]

This ruling could offer little more than cold comfort for Capay, presumed innocent but bearing physical and invisible scars from which few could ever hope to fully recover—and at the same time, depriving the public of its interest in having a serious case tried on its merits. And

one cannot help but question how such circumstances could have occurred in a Liberal democracy priding itself on its protection of fundamental human rights. And why, after a decade and multiple court challenges later, the practice of solitary confinement persists as a prevalent aspect of Canadian criminal justice practices, as well as in other Western countries.[10]

Jesper Ryberg correctly argues failing to account for penal theoretical issues risks rendering any discussions of using neurointerventions premature (Ryberg, 2021). This chapter is important as it will provide important context for discussions throughout this dissertation. In discussing our current practices of punishment, I think Capay's example fittingly illustrates the sorts of penal theoretical issues I am interested in here.

Punishment equivalence arguments suggest that if traditional punitive methods, such as imprisonment, are justifiable, at least in principle, neurointerventions could be comparable or superior alternatives. Such a discourse rests on the presupposition that the state has legitimate grounds for punishing crime offenders under specified circumstances.

Here, I generally accept that, at the highest level of ideal theorizing, the state is justified in punishing. However, this chapter falls in the realm of non-ideal theory and underscores the importance of considering practical limitations, empirical factors, and institutional context for any real-world application of biomedical procedures like neurointerventions. The chasm between theoretical punishment and practical implementation challenges punishment equivalence by introducing current practices as a comparative baseline.

More importantly, I hope to show the discourse around neurointerventions as a 'treatment for crime' may unintentionally simplify the complex nature of criminal behaviour. Shifting our focus to the neuroscience of ethics, I suggest the greatest benefits of modern neuroscience are not the tools they provide to forcibly intervene in the brain of disfavourable persons but those that provide growing insights into the true sources of crime, and the sorts of creatures we are and societies we inhabit and sustain.

In this sense, I argue analysis in Chapter 1 underscores strong prudential or practical reasons we have to be extremely cautious about a broad project for using neurointerventions in criminal justice practices and suggests a normative asymmetry between such interventions and conventional forms of punishment. It also prompts us to consider the broader biosocial context of crime, and its root cause may lie in places we might not expect.

An Irrational System of Punishment

Prevailing punishment practices in the United States and many Western countries can be reasonably characterized as irrational, considering any reasonable understanding or justification for punishment. They fail to effectively achieve valid penological objectives on any defensible theory of punishment. There are many reasons for this, but alongside many other theorists, I take the phenomenon of mass incarceration to be a predominant feature.

Punishment and Responding to Crime

Defining punishment proves elusive,[11] but I adopt an understanding of punishment as "a response to crime" that serves legitimate penological aims (Brooks, 2021, pp. 66-67; Sifferd, 2020). I accept neurointerventions could, *in principle*, be incorporated as part of a component of a larger "punishment package" or an integrated element of sentencing (Ryberg, 2018, pp. 135, 178).[12]

But to say we are justified in punishing, we must, as Pugh states, identify "how the institution of punishment itself is justified" (Pugh, 2018, p. 112). This involves exploring penal theories and punishment theories, which outline the purpose and methodology of punishment. The fact remains there is no consensus on "the most plausible ethical approach to punishment" (Ryberg, 2020, p. 190). In fact, many theorists also argue that a single ethical theory cannot wholly justify punishment (Brown, 2002; Sifferd, 2020, p. 296). It is true the reasons for punishment are manifold and include a mix of retributive (deontological, backward-looking) objectives, as well as consequentialist (forward-looking) objectives. Whether or not neurointerventions are an acceptable alternative to incarceration should be evaluated against these diverse and sometimes inconsistent objectives, which complicate the discourse significantly.[13]

But at the highest level of ideal theorizing, I do not think this poses significant obstacles for punishment equivalence arguments in a certain class of cases. There is no denying that there are individuals with the capacity to commit callous acts that inflict catastrophic harm, shocking the conscience and striking to the core of fundamental values of our communities. There is undoubtedly a core domain within which punishment may justifiably act to serve defensible ends—rape, murder, and torture, among others (DeGrazia, 2014, p. 364; R. Sparrow, 2014, p. 22). The extent to which the vast majority of those implicated in our current penal practices

would be seen to reflect this extreme class of case or the class of criminal conduct falling under a broad consensus of criminality—I suggest very few on either account.

But it is sufficient to say, theoretically, that the state is justified in responding to crime with punishment. Various theories of punishment could justify this, providing insights into ideas of desert, societal goals, censure, rehabilitation, and the forfeiture of certain rights upon committing a crime. However, the question is whether our existing penal practices, in the real world, effectively serve these theoretical justifications and goals. They do not.

The Birth of the Prison—Historical Context of Punishment Practices

Before we investigate the modern discourse surrounding neurointerventions in the criminal justice system, it is crucial to acknowledge that while such technologies are novel, conceptually, they are not without historical precedent. The evolution of punishment—from corporeal penalties to more subtle forms of psychological influence—has a rich historiography that provides valuable insights into our current debates.[14]

The trajectory of penal reforms, notably from the 18th century onward, reflects significant shifts not only in the methods of punishment but also in the underlying philosophical justifications. As we consider the introduction of neurointerventions today, it is instructive to recall that the penitentiary itself was once a novel technology. Initially conceived as a humane alternative to the brutal corporal punishments of earlier eras, prisons were designed to reform the mind rather than ravage the body. This reformative approach, culminating in the 'birth of the prison,' was underpinned by emerging Enlightenment ideals, which advocated for the dignity and potential improvement of all individuals, including persons convicted of crimes (Foucault, 1977). Sadly, what came to pass was often a far cry from these enlightened aspirations.

Historical analyses, such as Ted McCoy's (2012) study of Kingston Penitentiary, reveal that while these institutions aimed to embody progressive ideals, they frequently fell short, succumbing instead to practices that perpetuated harm and oppression. McCoy's detailed examination of the nineteenth-century reforms at Kingston Penitentiary demonstrates how the lofty goals of reform were consistently undermined by the realities of administrative corruption, inadequate funding, and societal indifference toward the plight of the incarcerated (McCoy, 2012)—as reflected in Capay's treatment over 100 years later within the progeny of the same corrections system.

In tracing the history of these institutions, we uncover a pattern where initial reform efforts, full of promise and philosophical vigour, gradually give way to the entrenchment of punitive practices that do little to rehabilitate or humanize those subjected to them. This cyclical failure not only highlights the resilience of punitive impulses over rehabilitative ones but also serves as a cautionary tale for modern reformers advocating for neurointerventions. As we venture into this new territory, it is imperative that we heed the lessons of the past to avoid replicating the same systemic failures that have long plagued our penal systems—which, as we will see, persist in pervasive forms.

A Crisis in Our Criminal Justice Practices

Hundreds of years removed from the 'birth of the prison', the current state of our penal institutions reveals a profound crisis that extends beyond mere description.

In recent decades, the rates of incarceration in Western countries have witnessed an astronomical and unprecedented surge. In Canada, the incarceration rate stands at approximately 104 per 100,000 (Canada, 2018; Malakieh, 2020).[15] The prison population in England & Wales quadrupled between 1900 and 2018, with a current rate of 167 prisoners per 100,000 (Berman & Dar, 2013, p. 4; Sturge & Tunnicliffe, 2021). The United States has experienced a quintupling of its incarceration rate in recent decades, with a staggering total prison population of 2.1 million individuals, equivalent to nearly 655 people per 100,000 (Garland, 2001; Ryberg, 2020, p. 189; Sawyer & Wagner, 2019). These figures mark an unprecedented phenomenon not only in the history of the United States but also in the history of liberal democracy (Garland, 2001; Ryberg, 2020, p. 189).

The annual cost of housing an inmate in Canada is estimated to be around $115,000, more than double Harvard's tuition cost (Canada, 2018)—foreseeably a fraction of the cost of programming rehabilitative programs addressing mental health or substance use programs. In the UK, it is estimated that the average cost to keep an inmate in prison exceeds £40,000 per year, and the total cost to taxpayers due to criminal reoffending ranges from £9.5 to £13 billion annually (Adebowale, 2010; Birks & Buyx, 2018, p. 133; Henrichson & Delaney, 2012, p. 6; Newton et al., 2019, p. 17). The United States spends an estimated $81 billion annually on its mass incarceration project (Wagner & Rabuy, 2017).[16]

Despite these astronomical costs, evidence suggests incarceration does little, if anything, to aid in rehabilitation or reduce the risk of reoffending. It fosters, contributes to, and even *causes* criminal behaviour (Brooks, 2021, p. 51; Cullen et al., 2011; Nagin et al., 2009; Ryberg, 2020, p. 191; Tonry, 2017). As supported by over 30 years of research and empirical studies, it is widely acknowledged that incarceration does not deter crime (Brooks, 2021 Ch. 2; Currie, 1989; Ryberg, 2020, p. 191; Sifferd, 2020, p. 304; Snodgrass et al., 2011; Tonry, 2008; 2011a, p. 58; 2017). A US study showed that 60% of released offenders reoffended, while another study found that two-thirds of released offenders were arrested for serious offences within three years (Brooks, 2021, p. 51; Doob & Webster, 2003, p. 51; Durose et al., 2014; Langan & Levin, 2002).[17]

Individuals entering the justice system often hail from backgrounds fraught with trauma, poverty, and marginalization. They are frequently plagued by mental health disorders, particularly Substance Use Disorder (SUD)—one of the most globally stigmatized health conditions (Rundle et al., 2021, p. 842).[18] These underlying conditions are exasperated and perpetuated by detrimental conditions within prisons, where prisoners endure conditions characterized by social exclusion, prolonged social isolation (particularly through practices like solitary confinement), overcrowding, inadequate healthcare, violence, and sexual assault (Baskin-Sommers & Fonteneau, 2016; Olivia Choy et al., 2018, p. 35; Focquaert et al., 2020, p. 141).

Upon release, these individuals often face stigmatization, social isolation, and the loss of personal and professional relationships, leading to secondary harms to their family members and communities and further harm to victims of crime (Brooks, 2021, p. 52; Olivia Choy et al., 2018, p. 36; Kleinig, 2012; Tonry, 2011a). Identifying this issue may point to issues of *overcriminalization*. Generally, too many acts are legislated to represent criminal acts when they should not be—for many reasons (Bradley, 2018 Ch 1; Husak, 2007, 2008; Larkin, 2013; Ryberg, 2018, p. 191; cited in Ryberg, 2020, p. 190). For example, drug possession for personal use exacerbates a public health crisis and has spurred a destructive cycle of imprisonment and release, leading to innumerable overdose deaths, despite desperate pleas for policy reform and evidence-informed harm reduction measures (Haley, 2020; Hendershot, 2020; Hrymak, 2018;

Mielau et al., 2021; Nadelmann & LaSalle, 2017; Spaniol et al., 2020; Strike & Watson, 2019; Watters, 2021).

In light of these challenges, penal theorists widely acknowledge that the current implementation of legal punishment falls short of its intended meaning and desired outcomes. It is regarded as ethically undesirable, driven by misguided policies, and even condemned as a "tragedy and national embarrassment" in the United States (Hardcastle, 2020, p. 163; Heffernan & Kleinig, 2000; Husak, 2008; Ryberg, 2020, p. 19; Tonry, 2017, p. 441; Nicole A Vincent et al., 2020a, p. 28). So, where have things gone so wrong? The issue comes down to this. Our current practices of punishment, as they exist in the real world, are irrational.

Confronting the Irrationality of Our Punitive Systems

What does it mean for a punishment system to be rational? It goes beyond justifying punishment based on individuals' preferred theory of justification. True rationality lies in implementing the system in a way that effectively *achieves* the intended penological objectives. A rational system is one not driven by emotion, intuition, or visceral instinct but draws sound reasoning, logic, and available scientific evidence to ensure the goals of the theory are met (Bedau & Kelly, 2019; Ryberg, 2020, p. 194; see also Tonry, 2011b).

In the context of a specific social or political system, this also requires punishment align with the underlying values, practices, and laws that govern the exercise of state power. We need not ground absolute moral claims about right or wrong; rather, we examine consistency and coherence. Moreover, because the state wields immense coercive power as the sole authority in administering punishment, poorly designed or implemented systems present a significant risk for the abuse of power and can have devastating consequences (Bedau, 1972).

In the face of prevailing conditions, can our institutions of justice be considered rational? It can not. It is more probable defensible penal theories will reveal the radical imperfections of our existing practices and explain why they fall short of the intended meaning and goals that would provide adequate justification (Heffernan & Kleinig, 2000).

For example, limiting retributivism, a dominant theory of punishment in the Anglo-American world, justifies punishment on the need to give offenders their just deserts, fit the punishment to the crime, and restore balance to the scales of justice (Frase, 2003; Matravers, 2018, p. 71; Morris, 1973; see discussion in Sifferd, 2020, pp. 297-300; Tonry, 2011b; Von

Hirsch, 2017). It enforces a substantial proportionality limitation to maintain moderation and avoid the risk of punishment becoming a mere instrument of retaliation or revenge concealed under the guise of legal authority (Bedau, 1978; Davis, 1992; Fingarette, 2013; Moore, 1987; Pereboom, 2009; 2018, p. 87; Pincoffs, 1977; Primoratz, 1997; Tonry, 2011b; Walker, 1991; White, 2011). Unfortunately, we will see such sentiments are a powerful and pervasive feature of our primitive neurobiology, which explains why this ideal has not been realized in practice.

Across various schools of thought in penal theory, including those with retributivist leanings, there is widespread agreement that offenders are often subjected to excessively harsh punishments that are disproportionate to the severity of their crimes (Frase, 2003; Matravers, 2018, p. 71; Morris, 1973; Sifferd, 2020, pp. 297-300; Tonry, 2011b; Von Hirsch, 2017). Such disproportionate sentencing is considered one of the primary contributors to the issue of overincarceration prevalent in many Western countries (Hirsch, 1996; Murphy, 1979; Roberts & Ashworth, 2016, p. 313; Ryberg, 2018, p. 191; Singer, 1979; Von Hirsch, 1992).

Consider communicative theories that view offenders as rational moral agents and propose that punishment should communicate censure while facilitating a self-reform process to enhance their moral character, resembling a type of 'secular penance' (Bennett, 2008; Boonin, 2008; Bublitz, 2018, pp. 312-313; Bülow, 2020; Duff, 1991, 2007; Duff & Duff, 2001; Duff & Hoskins, 2019; Markel, 2012; Matravers, 2000; Von Hirsch, 1996). It is difficult to see how exposing a lawbreaker to predictable risks within a setting characterized by social isolation, as well as physical and sexual assault, might inspire personal change. Contrarily, such surroundings can, and frequently do, inflict profound and often unalterable neurobiological harm to the rational, deliberative, and social capacities vital for fostering constructive moral discussions. Regrettably, a theory grounded on the notion of a state employing such measures to impart 'moral wisdom' rings hollow.

If these conditions communicate anything at all, it is a blatant disregard for prisoners' rights. This contradicts the goals of contemporary liberal theories of punishment that prioritize social justice, fair distribution of rights and obligations, diverse aims of punishment, and the paramount importance of human rights for persons, including those convicted of crimes (Bedau & Kelly, 2019; Dolinko, 1991; Duff, 1991, pp. 178-186; Goldman, 1982; Hoekema, 1986; Kant, 1797; Kaufman, 2008, pp. 45-49; Lewis, 1953; Moore, 1997; Morris, 1968; Murphy, 1973; Tadros, 2011; Ten, 1987; Von Hirsch, 1996; Wood, 2002).[19]

The same holds for rights forfeiture theories of punishment, which argue that punishment is justified on the grounds that individuals 'forfeit' certain rights when they commit specific criminal offences (Goldman, 1979; Morris, 1991, p. 68; Quinn, 1985, pp. 332-333; Simmons, 1994, p. 149; Wellman, 2009, 2012, 2017, 2020).[20] But these theories also stress persons cannot forfeit rights without exceptions and that individuals involved in criminal offences do not lose their status as "rights-bearing creatures" (Lippke, 2001, p. 87; Martin, 1993; Quinn, 1985, pp. 332-333). But in reality, prisoners do.

The undeniable truth is that we have failed to grasp vital moral lessons from a dark history. Despite the ongoing contemporary human rights revolutions and an increasingly scientific understanding of the destructive impacts of incarceration, we persist in disregarding the dignity and right-bearing status of prisoners. They are discarded as moral outcasts and warehoused by the millions in exclusionary and dehumanizing environments and subjected to treatment amounting to torture.

These violations further frustrate the ends of 'rehabilitation theory'[21] that states punishment should focus on reforming offenders and facilitating their transition from criminality to law-abiding citizenship, often as part of a broader social project acknowledging the eventual reintegration of incarcerated individuals into society (Brooks, 2021, pp. 66-67; Matravers, 2018, p. 73; McNeill, 2012; Raynor & Robinson, 2005; Sifferd, 2020). A system proficient in fostering rehabilitation would reduce crime—a goal that our current prison system fails to reach. A system of mass incarceration that inflicts foreseeable harm and communicates indifference proves costly, ineffective, and, ultimately, criminogenic. It promotes the occurrence of crime instead of mitigating it.

These are our current practices of punishment. In the real world, it is the 'baseline or comparison' for punishment equivalence. Where did things go wrong? The issues are complex, and so are the solutions. But I will revisit this issue shortly.

The Baseline Objection

Punishment equivalence arguments are grounded in the assumption that there is a justifiable foundation for the imposition of punishment. In theory, such a basis exists. However, our present-day practices of punishment are not morally defensible and deviate from this ideal.

This discrepancy raises a compelling objection to punishment equivalence, commonly called the 'baseline objection.'

Minimal Incarceration

Ethicist Thomas Douglas has put forth an influential punishment equivalence argument, questioning the intuition that valid consent is needed for administering neurointerventions to criminals under ideal situations (2014b, 2014c). Recognizing the inherent flaws in our existing punishment systems, Douglas adopts the concept of 'minimal incarceration as a benchmark for evaluation and comparison. He envisions a system where offenders experience significant restrictions on their freedom of movement and association yet face no more health and security risks than average free citizens. This system also takes reasonable measures to ensure offenders' access to political participation, legal representation, and education (2014b, p. 105). But reliance on idealized philosophical constructs, such as minimal incarceration, has attracted significant criticism. The most compelling line of argument is called the 'baseline objection.'

The Baseline for Comparison

There are those who challenge the defensibility of Douglas' ideal of 'minimal incarceration.' Some suggest that a fair punishment system should detain only the most dangerous offenders who are beyond the scope of psychological intervention. They argue that even minimal incarceration is unjustified or unnecessary if alternative methods can effectively achieve the aim of rehabilitation (Barn, 2019, p. 7; see also generally Pereboom, 2018).

Yet, even as these critiques challenge retributivist principles, they ought to simultaneously consider consequentialist considerations. An encompassing punishment system must strike a balance: it should honour individual rights while also efficiently deterring potential misdeeds and safeguarding society from further harm—even outside extreme cases involving the most dangerous offenders.[22]

However, even if we were to accept Douglas' vision of incarceration as an ideal to strive for, its realization remains unattainable in reality. This vision fails to accurately depict the current state of prisons, and its implementation in the near or distant future is highly improbable (Olivia Choy et al., 2018, p. 36).[23] But what are the implications of this for punishment equivalence?

Both incarceration and neurointerventions often infringe upon ethical principles such as autonomy, dignity, bodily integrity, and mental liberty (Canton, 2017; Focquaert et al., 2020,

pp. 142-134). But even were neurointerventions the 'lesser evil' of the two, it does not follow it is ethically permissible to administer them. According to Ryberg, regarding neurointerventions, their best attribute may be that they are 'the *least* evil'—but nonetheless, still be 'evil'—still being morally objectionable.[24] This alone does not provide enough justification to have confidence in their ability to effectively mitigate harm, especially in light of the significant concerns stemming from the irrationality of existing penal practices (Ryberg, 2020).

This leads to a significant concern that the implementation of neurointerventions under current conditions might set off what Matravers described as a "race to the bottom" (Matravers, 2018, p. 83). This phrase captures the potential for our standards to decline to the lowest acceptable ethical levels when new solutions are rapidly adopted without comprehensive evaluation. Neurointerventions, seen as a preferable alternative due to being less harsh than existing penal methods, might be accepted not because they fully meet ideal ethical standards but because they represent a marginal improvement over the *status quo*.

The "race" represents the collective and perhaps precipitous move toward adopting these interventions as stakeholders seek quick fixes rather than enduring solutions. This scenario often unfolds without a thorough consideration of the long-term consequences or the ethical dimensions necessary to genuinely reform penal practices. Vallentyne articulates this concern clearly, stating, "If (as I believe) our incarceration practices are typically impermissible, then my argument that consent to neurointerventions can be valid as an alternative to incarceration *does not apply to actual practice*. We would first have to change our incarceration practices to make them permissible" (Vallentyne, 2018b, p. 130 emphasis added).

The 'baseline objection has been widely discussed, and I adopt it with full force (Canton, 2017; Olivia Choy et al., 2018; Douglas et al., 2013; Focquaert et al., 2020, pp. 142-134; Kutcher, 2010a; Matravers, 2018; McMillan, 2014; Pugh, 2018; Ryberg, 2020; Sifferd, 2020; Vallentyne, 2018b, p. 130; Vanderzyl, 1994). In non-ideal theorizing, it generates compelling prudential—practical—reasons against the use of neurointerventions in many prospective applications, at least in the near future. However, for the remainder of this chapter, I aim to delve deeper into the complications that emerge when we concentrate on ideal theorizing.

Ideal Theorizing and What Matters Morally

As we have seen, most theorists, including Vallentyne, who put forth equivalency arguments, acknowledge the practical realities and exercise caution when addressing implementation concerns. In a way, this helps to establish a clear demarcation between the realm of ideal and non-ideal theory (Douglas, 2014c, p. 105; Ryberg, 2020; Vallentyne, 2018b).

As mentioned in the introduction, ideal theorizing is a predominant approach for punishment equivalence arguments. The express aim is to challenge intuitions that might otherwise distract from addressing issues in a constructive manner. So, setting aside concerns about our current practices and relying on a form of 'minimal incarceration' is not inherently objectionable. As Ryberg asserts, thoughts and arguments developed within hypothetical contexts "may often turn out to be applicable and illuminating in relation to other problems, some of which may be genuinely urgent" (2020, pp. 11-12). And I follow through many of these themes in the work that follows.

But when it comes to punishment equivalence arguments, there are growing concerns about the risks ideal theorizing might pose. Again, scholars caution against an extensive focus on such theorizing, suggesting that it may inadvertently "obscure some of the ethical aspects that truly matter" (McTernan, 2018b, p. 284). Extensive reliance on artificial philosophical constructs—such as minimal incarceration—may fail to address the worrisome and "potentially devastating real-world side effects that biomedical interventions may bestow upon individuals" (Focquaert et al., 2020, p. 142). I align myself with other scholars who entertain similar concerns (Bennett, 2018; Birks & Buyx, 2018, p. 135; McTernan, 2018b; Shaw, 2012; Vallentyne, 2018b, p. 138). Here, I identify two further.

First, despite the benefits of ideal theorizing in challenging problematic intuitions and traditional views, it presents a risk of inadvertently exacerbating issues within actual penal policies, especially when contemplating nonconsensual neurointerventions.

Neuroethics, as a specialized subset of applied ethics, plays a crucial role in shaping legislation, policy, and decisions surrounding neurotechnologies. It seeks to evaluate their effects on individuals and society at large, guiding their responsible and ethical usage. This is evident in the growing number of global organizations and institutions with this express mandate.

But while theorists engaged in ideal theorizing are generally explicit about the limitations of their approach (2014c, p. 105; Ryberg, 2020, p. 22; Vallentyne, 2018b, p. 130), the fact

remains that history teaches such cautions are not always heeded in practice. The lofty realms of ideal theory and the practical realities of penal policy in the messy world and the circumstances of the contemporary political arena are two different things. There is a significant gap to bridge. And it is hard to imagine an institution with a more troubled history of hastily embracing poorly considered policies than those of criminal punishment. To suggest otherwise or claim we have truly learned lessons from the past would require one to account for the current state of our penal practices.

I acknowledge it would be difficult to substantiate empirical claims on this basis. So, I simply highlight it is vital to exercise caution. If our goal is to positively influence actual penal policy, we need to acknowledge that our theoretical discourse unfolds within a setting characterized by misguided policies. It is vital to distinguish between ideal and non-ideal theorizing, coupled with transparency about our underlying theoretical assumptions.

The second concern I have is that while ideal theorizing has merit in identifying potential urgent issues, such as the use of emerging neurointerventions in penal practice, it risks narrowing our focus. As Ryberg suggests, ideal theorizing could bring 'genuinely urgent' matters to light (2020, pp. 11-12), but we must question if they represent the sole or most pressing issues in contemporary punishment practices. Crime is a complex social issue, and addressing it requires considering a diverse range of solutions beyond what ideal theorizing presents.

For our purposes, I believe we must confront challenging truths and remain receptive to the possibility that the root causes of crime, as well as the desired solutions, cannot be solely identified or addressed by only examining the synapses and nerve fibre tracts of individuals deemed unfavourable. If our genuine goal is to address crime comprehensively, the most promising solutions might lie in places we do not expect.

Objections: Critiquing Idealization Through Idealization

Before proceeding, it is important to acknowledge a further level of complexity introduced by critiquing the use of ideal theorizing within punishment equivalence arguments. This critique highlights the paradox of employing ideal models to challenge neurointerventions while demanding they align with a morally sound baseline. Questions arise about the feasibility of such a system, especially when addressing high-risk offenders, where prospects of rehabilitation are dim and there are serious risks of harm to others. Proposed paradigms focused solely on humane incapacitation might mirror the theoretical flaws of the idealized models we

examine. This scenario creates a fundamental paradox: advocating for an ideal system that may never be practical or achievable could trap us in the cycle of ideal theorizing we aim to critique. In essence, we critique PEAs for relying on ideal theorizing, yet we also propose a baseline that depends on similarly idealized, and ultimately unrealistic, visions of how penal systems should operate.[25]

There is merit to this concern. It is crucial to clarify that considering proposals for an idealized baseline for punishment is meant to challenge the punishment claim. This approach highlights the disparities between theoretical ideals and the practical realities of current penal systems, underscoring why the punishment claim is problematic. The goal is not to advocate for an impractical or utopian model but to critically examine how such ideals are presupposed in punishment equivalence arguments without sufficient regard for their feasibility or ethical implications.

This reflection enriches the overall concern—equally problematic idealized neurointerventions and an idealized baseline proposed by those such as Douglas—and reflects a general need to ground theoretical propositions in empirical realities. It encourages a deeper examination of the assumptions underlying current penal theories and aims to foster a dialogue that bridges theoretical aspirations with practical implementations.

We may lack a clear vision of what an ideal, just system of punishment would entail. But it is clear that our current systems do not meet such standards. My analysis focuses on identifying this gap, highlighting its significance for punishment equivalence arguments. This discrepancy demands attention, particularly when considering how our current practices have strayed and what steps are necessary to move towards a more just and humane system. Insights from neuroscience could also illuminate the barriers to achieving this goal—an issue to which we now turn.

Punishment and the Sort of Creatures We Are

Recall in the introduction of this dissertation, we distinguished between two branches of neuroethics: the *ethics of neuroscience* and the *neuroscience of ethics* (Roskies, 2002; Roskies, 2020a). The final part of this chapter focuses on the *neuroscience of ethics*. When it comes to the current crisis in our institutions of punishment, I think it is important to consider how advancements in neuroscience cast light into ‘the kind of creatures we are’ and ‘the social

structures that we inhabit and create' (Levy, 2007; Roskies, 2020a). We introduce certain themes here but will consider them in greater depth in the next Chapter.

The Dehumanizing Power of the Prison Environment

In 1971, a controlled experiment was conducted with 24 college students to investigate the effects of prison life. Researchers screened subjects to ensure they were deemed healthy and well-adjusted on measured psychological dimensions. They were randomly assigned to either prisoner or guard roles. Over six days in a simulated prison environment, the participants underwent a 'temporary but dramatic' transformation. The 'guards' began to subrogate the 'prisoners.' One prisoner even endured confinement in a makeshift solitary confinement room, resulting in a reported 'mental breakdown.' The cruelty escalated to the point that the experiment was prematurely terminated (Haney et al., 1973; Haney & Zimbardo, 1998; Musen & Zimbardo, 1992; Zimbardo, 2007; Zimbardo & Haney, 2020).

The widely known 'Stanford Prison Experiment,' a mainstay of first-year psychology courses worldwide, remains a subject of controversy and has faced valid criticisms. (Carnahan & McFarland, 2007; Le Texier, 2019). However, many researchers have addressed these concerns and argue that it still serves as a powerful demonstration of the impact of dehumanizing environments, such as the prison setting (Zimbardo & Haney, 2020). Leaving aside the debates surrounding the experiment, it is worthwhile to consider the thought-provoking questions posed by the researchers themselves: How did the inhumanity of the 'evil situation' overpower the humanity of the 'good' participants (Zimbardo, p. 62)?

The Sorts of Creatures We Are

Diogenes emphasized the importance of self-reflection and personal growth, stating that one can become one's own teacher by first acknowledging and reproaching within themselves the same faults they criticize in others (Laertius, 1853, pp. 256-259). What can neuroscience teach us about ourselves and the sorts of creatures we are when it comes to punishment? The hard truth we must face is that we are, by nature, innately myopic, vengeful creatures, hampered at every turn by evolutionary primitive, emotional responses that lie beneath the level of our conscious awareness.

Advancements in neuroscience, some of which we consider in the next chapter, continually unveil the inherent limitations of human rationality. This understanding departs from

folk-psychological notions of a mysterious and powerful 'unconscious' and is instead grounded in the insights offered by neuroscience into function and even the very architecture of our brains. Notably, the absence of vital connections mediating information between lower evolutionary primitive regions hampers our higher and distinctively rational human capacities. A range of heuristic and cognitive biases further clouds our global workspace of consciousness, obscuring our ability to fully grasp the inner workings of our minds (Bargh et al., 1996; Bateson et al., 2006; Davies, 2009; 2020, p. 339; Gazzaniga, 2000; Mitchell et al., 2011; Nettle et al., 2013; Pronin et al., 2008; Pronin & Ross, 2006; Uhlmann & Cohen, 2005; Wagner et al., 2012; Wilson, 2002).

In turn, our comprehension of the true origins of our decisions, actions, and reactions remains substantially limited. We remain "strangers to ourselves"—a sentiment that may have been all too apparent to the 24 college students involved in the Stanford Prison Experiment. (Bechtel, 2007; Davies, 2009, 2020; Panksepp, 2004, 2012; Wilson, 2002). And this may be "difficult to accept because, naturally, it is alien to our experience, *but it is true*" (cited in Kahneman, 2011, p. 52; Ryberg, 2020, p. 90).

Within the lower basal regions of our primitive brains, a multitude of robust affective responses emanate as odd vapours, propelling and distorting our views of deserved punishment. Research conducted in the realms of neuroscience and evolutionary psychology establishes such intuitions to be intricately intertwined with our evolutionary history, endowing us with adaptive intuitions and corresponding neural mechanisms aimed at promoting cooperation and addressing individuals who violate social norms. These intuitions and their neural underpinnings were pivotal to ensuring survival in primitive small-scale cooperative societies (Cushman, 2015; Jean-Richard-Dit-Bressel et al., 2018; Mikhail, 2011).

Over millions of years of evolutionary history, these neural mechanisms and supervenient social institutions were finely calibrated to address issues in primitive, small-scale, closely-knit societies facing immediate issues. Generally, this required straightforward causal attributions of blame and norm violation in small groups facing immediate problems of survival. This is a reflected form of 'common sense morality,' which is myopic, evolutionary primitive, and leads to a form of 'moral tribalism' (Greene, 2013; Persson & Savulescu, 2011b, 2012, 2015).

As to the social structures we have created, it is widely held that human societies have evolved to facilitate such intuitions about punishment and reflect the positive expression of our

moral intuitions (Mikhail, 2009, 2011). Both in primitive and contemporary societies, social structures and corresponding institutions 'exploit' such evolved intuitions to achieve cooperation, and "punishment can only be understood as an interaction between biology, culture, and institutions" (Cushman, 2015, p. 130).

Over the course of the past century and millennia, significant transformations have taken place in various domains that have fundamentally altered the conditions of human existence. But the sorts of creatures we are have not—at least not in any meaningful *neurobiological* sense. It is said this leaves us ill-equipped to navigate the complex issues we face in large-scale contemporary societies (Greene, 2013, p. 61; Persson & Savulescu, 2012, pp. 39-40). As Ingmar Persson and Julian Savulescu put it, we are 'unfit for the future' (Persson & Savulescu, 2012).[26]

Our evolutionary history was premised on primitive notions and simplified beliefs that prioritized physical causes over complex mental states and a comprehensive examination of social and causal history (Persson & Savulescu, 2012, pp. 22-23). On this basis, it is unsurprising humans possess strong intuitions towards retributive punishment, focusing on intent and the outcome of wrongdoing, motivated by a form of 'moral outrage' triggering affective responses notoriously susceptible to indirect subversions, heuristic and cognitive biases (Bastian et al., 2013; Carlsmith, 2008; Carlsmith et al., 2002; Cushman, 2008, 2013; Cushman, 2015; Darley et al., 2000; Fincham & Roberts, 1985; Haidt, 2001; Mikula et al., 1998; Robinson, 2019; Tetlock et al., 2000; Weiner, 1995).

Humans are highly sensitive to signals of group membership and intuitively disposed to favour in-group members over out-group members (Greene, 2013, p. 61; Persson & Savulescu, 2012, pp. 39-40). For example, tribalism, beneficial for small societies in resource protection, manifests as 'similarity' or 'affinity' bias, favouring those close or similar to us, and 'outgroup dehumanization,' perceiving norm violators or the 'deviant other' as similar to objects or animals lacking uniquely human features (Bandura, 1999; Bandura et al., 1996; Čehajić et al., 2009; Eagleson et al., 2000; Haslam, 2006; Mikula et al., 1998; Tetlock et al., 2000; Vasiljevic & Viki, 2013; Viki et al., 2012).[27]

According to Bruce Waller, fueled by these powerful forces, our reliance on a 'stubborn system of moral responsibility' forms the basis for flawed institutional structures (Waller, 2015).

The misguided responses to crime stem from our overconfidence in the capacity of human reason and powerful retributivist sentiments. However, for Waller and others, the problem with retributivist theories is that the notion of desert is not grounded in rationality or reason but in powerful emotions that motivate veiled tendencies towards revenge and vengeance or a 'strike back desire' (Pereboom, 2009; 2018, p. 87; Waller, 2015 Ch 11). These are the sorts of creatures we are and, as it stands, the sorts of social institutions we inhabit and continue to sustain.

The Root Cause of Crime

In the past fifty years, the prevailing theory of punishment in the Anglo-American world has been retributivism— 'eye for an eye' and 'strike back' (Matravers, 2018, p. 71; Tonry, 2011b; Von Hirsch, 2017; Whitehead & Chandler, 2018). However, cultures that exhibit strong retributive tendencies often coincide with high levels of social inequality, leading to increased disparities and harsh treatment of marginalized individuals in society (Waller, 2015 Ch 11). Why is this the case?

Individual Moral Ills and The Myriad Causes

At the outset, I reject the revisionist view that people cannot act freely or be morally responsible for their actions (Greene & Cohen, 2004; Pereboom, 2009, 2014, 2018; Sapolsky, 2004; Shaw et al., 2019). Acknowledging the nuances, there are instances, such as those involving repeat offenders with cognitive impairments, where the impetus for punishment pivots from retributive to consequentialist reasons. In these cases, the societal imperative for safety and order might warrant incarceration or even involuntary hospitalization. This highlights the need for a layered approach to punishment that equally factors in moral culpability and broader societal impacts.

Yet, the grip of retributivism on Western penal policy remains undeniable. But it is based on a simplistic view of criminal behaviour.[28] I pause to recognize that retributive instincts are not inherently flawed. Altruistic punishment, where individuals incur personal costs to enforce social norms, demonstrates that retributive actions can play a critical role in evolved psychology within social species. These instincts might compel individuals to protect defenceless victims rather than turn aside with indifference, reflecting a complex interplay between evolved moral responses and social cohesion.

Moreover, while retributive instincts are often dismissed as primitive, this perspective overlooks their role in maintaining order and justice within society. Retributive justice is not just a reflex but a deeply ingrained part of our social fabric that reinforces norms and deters antisocial behaviour. We create sophisticated legal institutions that channel these instincts through processes designed to temper immediate reactions, allowing for a more measured and fair application of justice that aligns with modern values of equity and human rights.[29]

However, the primary concern is the degree to which retributive justice detracts from, distracts, or overwhelms other relevant considerations that are expressly endorsed and prized in Western liberal societies. These considerations include the rehabilitation of offenders, the prevention of crime through social reform, and the protection of human rights, which often require more nuanced and restorative approaches than those provided by mere retribution.

Understanding the sources of crime requires considering the complexities of human behaviour in advanced human societies, which are vastly different from those premised on primitive reactions, intuitions, and cooperative behaviours tailored to address the conditions of small, tightly-knit groups of our early tribal ancestors. In this light, while not dismissing the instinctive draw of retributive justice, we must also critically examine and potentially expand our penal practices to better reflect the sophisticated and diverse needs of modern, pluralistic societies.

Stemming in large part from powerful retributivist sentiments, and straightforward causal attributions, too often, practices of punishment are premised on the view that "humanity's moral ills [are] a result of individual moral deficits" (de Melo-Martin & Salles, 2015, p. 6; Ryberg, 2020, p. 15). And understanding our evolutionary history in this manner, it is not difficult to see why this is the case. But as Dostoyevsky observes, "The causes of human actions are usually immeasurably more complex and varied than our subsequent explanations of them" (Dostoyevsky, 1995 [1869]). Crime cannot be adequately understood by isolating the actions of an offender in a particular moment—much less their humanity, worth, or future prospects. As Coppola explains, "[o]ffending behaviour, like any form of social behaviour, is much more complex and better understood within a framework that takes account of the complex interplay of dynamic biological, psychological, and social factors" (Coppola, 2018, p. 13). Understanding this interplay, I think, serves as a useful tool for identifying the problem and a broader range of possible solutions.

To ensure a rational system of punishment, we need to "dig deeper than philosophical arguments" and acknowledge that many of our commitments to standard views of moral responsibility are more robust than arguments in its favour (Waller, 2015, p. 6). The most problematic effect of powerful primitive emotions is that they lead to our inability to identify and appreciate the moral significance of the "myriad of hidden causes of peoples' behaviour" (Waller, 2015, pp. 111, 119)—and, in turn, appropriate reactions to address undesirable behaviours in a constructive, evidence-based, and humane manner.

As Ryberg correctly notes, the idea of an association between criminal behaviour and the brain's functioning is far from new (Ryberg, 2020, p. 3). Neither is an appreciation of the devastating impact of social factors, such as poor socioeconomic status, childhood abuse and neglect, and intergenerational trauma, on the contribution to risk factors associated with future offending. However, remarkable advancements on the frontiers of neuroscience continue to provide profound insight into just how damaging these factors are in the development and even the structure and function of the brain.

The Brain and the Bonds that Unite Humankind

A significant focus of theorizing for the balance of this dissertation, particularly the next chapter, involves recognizing that the brain is a relational and social organ (Fuchs, 2004, 2008, 2011; Fuchs & Schlimme, 2009; Glannon, 2009). It assumes privileged status in our mental lives through its vital role in integrating, mediating, and scaffolding crucial aspects of our actions, behaviours, decisions, and capacity for critical and moral reasoning across embedded and extended domains (Clausen & Levy, 2015, p. 35; Kalis et al., 2008, p. 21). A natural corollary is the profound importance of healthy social connections in developing and sustaining healthy affective and cognitive functioning and adaptive social behaviours (Deckers, 2014; Levenson, 1999; Lieberman, 2013; Moors et al., 2013; Siegel, 2012). These—the 'bonds that unite humankind'—not only shape foundational regulatory structures in the brain but the neural connections that allow us to construct representations of reality and formulate a coherent view of ourselves, the world, and our place in it: "interpersonal experiences directly influence how we mentally construct reality" (Siegel, 2012, p. 9).

Conversely, the destruction of those bonds has a profound impact. This impact is particularly profound during the formative years of youth and is associated with deterioration in

brain regions associated with poor moral and social decision-making and maladaptive behaviours, such as antisocial conduct (Coppola, 2018, p. 5; Cozolino, 2014; D'Angiulli et al., 2008; Fox et al., 2010; Gillespie et al., 2017; Hart & Rubia, 2012; Kessler et al., 1997, pp. 460-461; Kishiyama et al., 2009; Koenigs et al., 2007; Kolk & Fisler, 1994; Leutgeb et al., 2016; Ma et al., 2011; Ostovar, 2009; Piotrowska et al., 2015; Raine, 2008; Siegel, 2012, p. 22; Sobhani & Bechara, 2011; Stevens et al., 2009; Weller et al., 2007).

Understanding the sources of crime, proportionate punishment, and seeking evidence-based solutions requires an appreciation of the contributions of systemic factors that encompass low social and economic status, systematic disadvantage, mental illness, homelessness, educational inequity, abuse, and Substance Use Disorder.[30] For example, research suggests that 50% of incarcerated offenders in the United States have mental health-related problems (Ginsberg & Lindefors, 2012; James & Glaze, 2006; Ryberg, 2020, p. 7). In many, if not most, cases, these underlying systemic issues stem from intergenerational issues that contribute significantly to the adverse circumstances faced by individuals. The cruel irony lies in the profound and lasting harm that often drives individuals into the criminal justice system, only to be placed in a dehumanizing environment characterized by the very elements that caused their suffering.

Extensive psychological and neuroscientific evidence highlights the detrimental effects of the prison environment. It can worsen or give rise to various dysfunctions, including affective disturbances, depression, anxiety, impaired self-regulation, persistent anger and rage, psychosis, diminished self-esteem, feelings of rejection and humiliation, and an increased risk of suicide. The exclusionary and restrictive conditions in prisons exacerbate a range of affective, cognitive, and behavioural deficits, while "enhanced social isolation and reduced physical contact contribute to and reinforce problematic neurobiological patterns" (Baskin-Sommers & Fonteneau, 2016; Olivia Choy et al., 2018, p. 36; Denno, 2016; DeVeaux, 2013; Dillon, 2018; Gallagher, 2014; Gilligan, 2000; Jacobs, 2016; Matravers, 2018, p. 83; Smith, 2015).

Viewed from a long-term perspective, this situation contributes to an overall increase in human suffering, including for victims of crime. Others suffer secondary harms, perpetuating intergenerational cycles and adversely affecting families and communities across space and time. The prison population continues to swell. Cruelty breeds more cruelty. Trauma begets further trauma. This perpetuates devastating intergenerational harms visiting their harshest consequences

against individuals who are already among the lowest in society. It leads to astronomical financial costs, widespread human suffering, and, in its wake, countless wasted lives.

Final Thoughts—Trends Towards Reform

So as to not end on a pessimistic note, I would note some promising developments that warrant attention. In recent times, there has been a shift in penal theoretical discourse towards Nordic penal systems, often praised as 'two-star hotels' due to their emphasis on comprehensive rehabilitative support, public protection, and reintegration. These systems not only exhibit lower incarceration rates but also prioritize more humane prison conditions, representing a remarkable departure from the prevailing punitive trend in Western penal practices (Barker, 2017; Davidson et al., 2000; Davidson & McEwen, 2012; Garland, 2001; Maier, 2020, p. 383; Nelken, 2009; Pereboom, 2018, p. 95; Pratt, 2007a, 2007b; Shammas, 2014; Smith & Ugelvik, 2017; Ugelvik & Dullum, 2012).[31]

Another promising development is reflected in a recent 'therapeutic jurisprudence movement', which aims to further rehabilitative and communicative ends of sentencing through justice responses based on more "rational, effective, and humane grounds" (Coppola, 2018, p. 9; Hardcastle, 2020, p. 162; Marlowe, 2021; Matravers, 2018, p. 72; Matusow et al., 2013; Sifferd, 2020, p. 309; Tonry, 2011b). This is reflected in the emergence of drug treatment courts across countries such as Canada, the UK and the US, specifically designed to target and address SUDs and mental health issues (Matravers, 2018, p. 72; Mitchell et al., 2012; Wolff et al., 2011).

Interestingly, these programs utilize novel psychopharmacological interventions such as topiramate and methadone maintenance treatment (MMT)—correctly classified as 'neurointerventions' High-quality randomized control trials have demonstrated a strong association between MMT and a reduction in drug use relapse. Preliminary data also suggests the effectiveness of these interventions in assisting with rehabilitation (Eley et al., 2002; Focquaert et al., 2020, p. 131; Hall & Carter, 2013; Hough et al., 2003) (Bales & Piquero, 2012; Focquaert et al., 2020, p. 131; Hall & Carter, 2013; Hardcastle, 2020, p. 159; MacKenzie, 2006; Mears, 2012; Stevens et al., 2005; White et al., 2012).

But the unfortunate reality is that our stubborn system of moral responsibility persists. Beyond these promising trends, I think what the future holds for our practices of punishment

remains to be seen. A full understanding of the problems we face and perhaps the possible solutions fall outside the realm of ideal philosophical theorizing—and most certainly, ideal theorizing relying on fictional constructs such as 'minimal incarceration.'

If solutions exist to the problem of crime, they will ultimately be realized in the real world and fall within the realm of public policy and law. Much more could be said about this, extending far beyond the scope of this dissertation. I simply note that I believe our potential for progress lies in our ability to better know ourselves and continue such an endeavour with compassion and humanity. Remarkable advancements in neuroscience are a powerful extension of Enlightenment ideals and the belief in humanity's ability to transcend natural limitations. But at present, I believe the most powerful tools they yield to address immediate issues are not those to forcibly intervene in the human brain but to better understand the true causes of criminal behaviour and, more importantly, ourselves.

Conclusion

Equivalency arguments in punishment rely on the assumption that we have the right to punish. At a theoretical level, these arguments can draw upon various penal justifications and use ideal theorizing to challenge problematic intuitions related to the use of neurointerventions in criminal justice. The success of such arguments as a whole is the focus of subsequent discussions, which raises important ethical issues.

But descending from the realm of ideal theorizing, it is crucial to recognize that crime is a complex social issue in the actual world in which we live. This chapter has illustrated that in the realm of non-ideal theory, there are strong prudential reasons to exercise extreme caution in rushing to utilize novel neurointerventions in contemporary justice practices. Our current punishment institutions are in crisis, typified by archaic practices that perpetuate cycles of harm and inflict untold suffering on those most deserving of sympathetic consideration.

Taken alongside a dark history of mistreatment of prisoners, a contemporary crisis illustrates we have not learned from past lessons and equipped with powerful tools for neuromodulations—which we will see may foreseeably pose devastating risks to safety—a deeper acknowledgement of our primitive, myopic nature generates strong practical and prudential reasons against adopting a comprehensive regime for implementing neurointerventions in our institutions of punishment. This poses what I see to be devastating

objections to the punishment claim, assuming the the aim of equivalency argument is to offer any practical guidance outside the lofty realms of ideal theorizing.

And among the host of issues we stand to address throughout the balance of the 21st century, as cutting-edge technologies intersect with criminal justice practices, we must not lose sight of an important question: how has the inhumanity of our current penal practices overcome humanity to which our Western liberal traditions ascribe?

Perhaps it is within our power to build a society that is less brutal, less fearful, and more humane. However, our ability to effectively evaluate the promises and pitfalls of novel technologies rests on our readiness to embrace their potential for self-discovery. And this demands that we remain open to the idea that these tools can empower us to confront and examine the faults and deficiencies within ourselves, even as we aim to condemn them in others.

2

Empirical Assumptions and Safety and Efficacy in Neurointerventions

What We Image and What We Imagine

Introduction

In the mid-20th century, the lobotomy emerged as an astonishing "miracle cure" for an array of mental disorders. Physicians, led by Portuguese neurologist António Egas Moniz, who received a Nobel Prize in Physiology or Medicine in 1949, cut into patients' brains to sever crucial neural pathways. This procedure, praised for its innovative approach, often resulted in severe, irreversible damages. In America, a prominent figure of this era, Dr Walter Freeman, evangelized the practice, performing thousands of lobotomies with a nonchalant attitude, overlooking the human costs that came with it (El-Hai, 2005; Johnson, 2014).

Regrettably, this controversial medical intervention was not limited to mental hospitals; it found its way into the criminal justice system. In a bid to control aggressive behaviour or "cure" supposed mental ailments, lobotomies were performed on imprisoned individuals across the United States (Kulynych, 2007).

In our history, the lobotomy stands as a stark reminder of a disconcerting era where the pursuit of scientific advancement overshadowed the ethical implications within the criminal justice system. It demonstrates how even well-intentioned innovations can lead to profound harm when we neglect the bounds of our understanding and the potential dangers that come with them. As we journey forth into new frontiers of neuroscience and criminology, it's crucial that we proceed with a blend of audacity and humility—daring to explore yet cautious of the limits of our knowledge and the possible perils they carry. This balance will be essential in ensuring we do not unwittingly echo the mistakes of our past.

Between Dystopian Visions and Promises of Progress

There are very few advancements during the 21st century that have captured the imagination as much as those of modern neuroscience. On the one hand, there are utopian visions

of progress and human perfection. On the other are, dystopian disasters immortalized in popular fiction. Randle McMurphy's unwavering resistance against the dehumanizing authority of a mental institution and harrowing lobotomization (Forman, 1970; Kesey, 2007). The chilling saga of experimental behaviour modification portrayed in Burgess' a *Clockwork Orange* (Burgess, 1962). A *Brave New World*, where citizens are controlled through advanced technology, manipulated reproduction, and a drug-induced illusion of contentment (Huxley, 2021 [1932]). Orwellian visions of pervasive surveillance technologies, manipulating the very fabric of reality, eradicating personal autonomy and subjugating the human mind to maintain absolute control (Orwell, 1977 [1949]).

In light of contemporary advancements, one might be surprised to learn that, as we will see, at least *some* such imaginable interventions can no longer "be relegated to a distant dystopian future" (Stefano, 2021, p. 211). But when examining the potential of contemporary neuroscience in addressing the complexities of crime, Judy Illes and Eric Racine aptly emphasize the importance of differentiating between our current visions and imaginative speculation, stating the need to "untangle what we image from what we imagine" (Illes & Racine, 2005, p. 12). By carefully discerning between these perspectives, we can gain a more nuanced understanding of the possibilities and limitations of neuroscience within the realm of criminal justice.

In reality, between utopian aspirations and dystopian visions, our ethical theorizing necessarily falls somewhere in between. In the face of rapid advancements, our ability to respond to and develop theories may lag behind, for as the inquiry persists, we must acknowledge that "we don't know what we don't know, but we should prepare for the answers we get" (Farahany & Ramos, 2020, p. 148).

In this chapter, I examine the second feature of punishment equivalence arguments, which involves a class of 'empirical assumptions'—generally related to 'safety and efficacy.' I discuss the issue of the "neuroscientific turn" in contemporary neuroscience and its implications. Emphasizing the intricate nature of the human brain, I present key facts and principles related to the safety and effectiveness of neurointerventions. I explore recent advancements in understanding the neural foundations of human morality. Additionally, I address the challenges involved in modifying human morality through existing and potential future neurointerventions. As with Chapter 1, I argue the current state of our understanding of neuroscience and the

limitations of contemporary neurointerventions pose an insurmountable challenge to equivalency arguments and the prospect they provide meaningful guidance to ethical issues.

I argue this supports a normative asymmetry between neurointerventions and conventional forms of punishment. However, I identify preliminary concerns about how these considerations raise conceptual and theoretical issues even in the realm of ideal theorizing—which underscores issues we will explore further in later chapters. I conclude by highlighting the importance of anticipatory ethics and the need for a comprehensive precautionary principle in the application of new neurotechnologies in criminal justice.

As a final note, this dissertation ventures into the sphere of applied philosophy and ethics, intersecting with natural sciences within a framework of established theories, laws, and foundational assumptions. As such, his chapter is a preliminary dive into contemporary neuroscience, serving as a critical lens through which we scrutinize these core principles that later chapters will build upon. The intention is to ignite discourse around both empirical and conceptual concerns, acknowledging the intricate layers within our discipline, especially in relation to human rationality, freedom, and the potential recognition of rights over the 'mind.'

In-Principle Differences

Punishment equivalency arguments suggest that if we accept the justification for conventional punishments for criminal offenders, we should also consider the possibility, *in principle*, of utilizing novel neurotechnologies as viable alternatives.[32] The arguments prevent the bias of presuming that they can be 'ruled out from the beginning' or 'dismissed out of hand' (Douglas, 2014c, p. 120; Greely, 2008, p. 1134; Ryberg, 2018, p. 180).

In this sense, punishment equivalence arguments claim to show neurointerventions and conventional forms of punishment are equivalent *assuming certain conditions hold*. I take the 'in principle' constraint to trace these conditions and, by extension, the range of assumptions required for these arguments to offer meaningful guidance in real-world situations.

An 'in principle' difference denotes a theoretical, rather than practical, distinction. This approach represents a prototypical form of 'ideal theorizing.' The aim is to construct a conceptual framework to comprehend ethical ideals and principles and acknowledges that "progress on the penultimate questions need not wait for solutions to the ultimate ones" while evaluating the

potential for creating a "reasonable pattern of moral arguments" (BonJour, 1998; Feinberg, 1987, p. 18; Rawls, 1971, p. 245; A. J. Simmons, 2010; Wolff, 2011).

This method reflects an analytical strategy introduced by Neil Levy (2007) in his pivotal book, *Neuroethics: Challenges for the 21st Century*. Levy elucidates 'in principle' differences in the sphere of neurointerventions as enduring disparities that subsist irrespective of the degree of technological advancement or the "political and social context" of their development and application (Levy, 2007, p. 73).

The 'in principle claim' is closely related to Levy's 'parity principle,' which we will explore in the subsequent chapter. His reasoning continues that *if* we assume neurointerventions are safe and effective *then* we should treat them as "ethically on par" unless we can identify ethically relevant differences in their *effects* (Levy, 2007, p. 62).

The 'in principle' assertion establishes a categorical differentiation that shapes the 'parity principle' and encourages equal scrutiny of mental states, regardless of their roots—whether they emerge from conventional methods or are facilitated by innovative neurotechnologies. In many ways, the two go 'hand in hand.'

But here, I address the 'in principle' aspect of punishment equivalence arguments separately. The reason for doing so is the significant importance of this particular aspect in the overall argument. It is possible punishment equivalence arguments may proceed to focus on an 'in principle' approach and still purport success in achieving their goal—challenging questionable intuitions. I recognize concerns about shifting the objective or parameters of the discussion. But initiating the exercise within the sphere of ideal theorizing through a search for 'in principle' differences yields significant gains that I believe cannot be overstated.

Similar concerns emerge when discussing idealized notions of "minimal incarceration," as explored in the previous chapter. Idealizing the safety or effectiveness of neurointerventions may divert focus from potentially severe real-world consequences while veiling ethically pertinent considerations. As I hope to show, the practical effect of this theorizing would set aside a host of contingent considerations and adopt a range of empirical assumptions, which would risk relegating ideal theorizing into the realm of science fiction. This would ultimately foreclose any ethnically responsible discussions of ambitious, practical application—particularly in criminal justice practices. But even within the realm of *ideal theorizing*, failing to account for those assumptions presents conceptual challenges and stifles productive theorizing.

Indeed, in reality, the concepts of ideal and non-ideal theory exist on a spectrum rather than being distinct and separate entities. It is important to acknowledge that the "in-principle" constraint and the range of corresponding assumptions can significantly differ in strength and application across various arguments. For example, in advancing punishment equivalence arguments, Lippert-Rasmussen presents a strong definition of an 'in-principle difference' as grounded in 'definitional properties'—perhaps representing *a priori* differences and corresponding considerations (Ayer, 1971; Kripke, 1980; Lippert-Rasmussen, 2018).

In similar theorizing, Levy appears to ground discussions on 'reasonably envisaged neurointerventions' and 'features of the world' so that the properties of neurointerventions could serve as a useful 'heuristic' for identifying ethical issues (Levy, 2020, p. 46).[33] He emphasizes the importance of accounting for contingent empirical considerations as part of an 'all-things-considered' judgement about use in real-world circumstances (Levy, 2007, p. 72).[34] I believe that Levy's empirical perspective is the most productive approach, and it will reflect the manner in which this dissertation addresses subsequent issues.

When we subject assumptions to scrutiny through empirical evidence, it offers a more effective approach to challenging intuitions and advancing ethical theorizing. This process enables us to ensure internal consistency, external coherence, simplicity, and explanatory power within our ethical frameworks. By grounding our discussions in empirical evidence, we can enhance the robustness and reliability of our ethical theories, making them more comprehensive and applicable to real-world contexts. Theoretical discussions centred exclusively on abstract logical possibilities, without considering their practical feasibility, all too often lead to conceptual issues, as I hope to show in the next chapter.

Exploring the Boundaries of Knowledge: The Unveiling of Brain and Mind in Contemporary Neuroscience

Neuroscience is a multidisciplinary field that studies the brain[35] and its connection to our mental life, exploring how this complex organ gives rise to human cognition, emotion, morality, consciousness, and other key features that make us uniquely human (Bear et al., 2020; Dowling, 1998; Felten et al., 2015; Glickstein, 2014; Kandel, 2013, p. 5; Squire, 2012).

The social sciences and humanities have increasingly delved into the role of the brain in various social and cultural phenomena. This interdisciplinary approach, combining empirical and

conceptual methods, seeks to gain clarity on the intricate relationship between the mind and mental life—fundamental assumptions about concepts like the self, consciousness, life, death, and our relationships with others (Farahany & Ramos, 2020, p. 150; Koroshetz et al., 2020, p. 145; Leefmann & Hildt, 2018, p. 14; Pardo, 2014, p. xv).[36]

Neurohype, Neuromania, and the Neuroscientific Turn

Since the 21st century, neuroscience has experienced remarkable growth driven by advancements in brain imaging technologies. These technologies enable the measurement and examination of brain activity, shedding light on the role of neurons and neural circuits in cognitive, affective, and volitional functions. This includes novel forms of structural and functional imaging.

Early structural imaging methods, such as computed tomography (CT), have been replaced by more advanced technologies like magnetic resonance imaging (MRI) and diffusion tensor imaging (DTI) (Bruno et al., 2011; Chen et al., 2016; Ciaramelli et al., 2007; Kretschmann, 2004; Lin et al., 2008). Functional imaging techniques—which provide temporal information about neural activity during cognitive tasks—have also been refined. For example, electroencephalography (EEG) and magnetoencephalography (MEG), which involve measuring electrical brain activity, have been augmented with advanced technologies tracking metabolic or neurovascular activity such as positron emission tomography (PET), single-photon emission computed tomography (SPECT), and most notably, functional magnetic resonance imaging (fMRI) (Ciaramelli et al., 2007; Gardner et al., 2006; Kretschmann, 2004; Lin et al., 2008; Lu & Yuan, 2015).

Functional imaging methods, particularly functional magnetic resonance imaging (fMRI), provide significant benefits through their ability to record brain activity. They achieve this by tracking alterations in blood flow and oxygen levels in the brain as subjects interact with their surroundings. This technology not only facilitates the observation of changes over time but also allows for the detailed mapping of spatial distribution of activity. Such capabilities enable the study of both localized activations and diffuse patterns of activity across the whole brain, making fMRI an indispensable tool in a wide range of fascinating neuroscientific studies.

For example, fMRI used to identify individuals' preferences for presented objects (Hosseini et al., 2011)[37] and aid communication with people with disorders of consciousness

(DOCs) (Abbott & Peck, 2016; Campbell et al., 2020; Owen, 2013; Owen et al., 2006), reveal dispositions to implicit bias (Chekroud et al., 2014; Greenwald et al., 2002; Korn et al., 2012; Phelps et al., 2000), detect neural correlates of deception and concealment—leading to a technology known as fMRI lie detection (Farah et al., 2014; Hsu et al., 2019; Joseph, 2008),[38] and significantly for our purposes, locate neural mechanisms associated with moral judgments (Anderson et al., 1999; Fede & Kiehl, 2020; Greene, 2013; Greene & Haidt, 2002; J. D. Greene et al., 2001; Koenigs et al., 2007; Tassy et al., 2012).

These advancements have certainly stirred the imagination and have coincided with what is known as a "neuroscientific turn" (Leefmann & Hildt, 2018). We have become our "neurochemical selves" (Rose, 2007), witnessing a cultural shift towards a "neuro-society" (Leefmann & Hildt, 2018, p. 16) fueled and accompanied by a phenomenon known as "neuro hype" (Lilienfeld et al., 2018). This trend is evident in the proliferation of "neuro" subdisciplines—neuroeconomics, neuroeducation, neuromarketing, neuropsychology, neurohistory, neuroaesthetics, neurotheology, neuroeducation, and for our purposes 'neuroethics' and 'neurolaw' (Battro et al., 2008; Glimcher & Fehr, 2013; Lee et al., 2007; Loewenstein et al., 2008; Stanton et al., 2016; Ulman et al., 2015). However, this phenomenon has led to two problems which we must be mindful of in the discussions that follow.

First, fueled by colourful images of the brain, neuroscience has what has been described as a 'seductive allure' and promotes overblown and misleading claims that researchers and media portray to the public (Lilienfeld et al., 2018, p. 243; Racine et al., 2006; Racine et al., 2017; Roskies, 2022; Satel & Lilienfeld, 2013; Schick, 2005; Tallis, 2016, p. 73; Tovino, 2007; Vidal, 2018; Weisberg et al., 2008). As Neil Levy explains: "[t]he aura of prestige and objectivity which surrounds science generally is perhaps even stronger about the science of the mind at its cutting edge" (Levy, 2007, p. 144). It is worth noting that some recent studies challenge this view by showing that neuroimages might not have as much effect as once (Bennett & McLaughlin, 2023).

Notwithstanding, claims made by neuroscientists regarding the accessibility, measurability, and predictability of the mind are met with skepticism for valid reasons.[39] Brain imaging techniques like PET, MRI, and fMRI provide visualizations of statistical analyses rather than capturing events and processes at the neuronal level. They are arguably better understood as

"scientific constructs" than actual brain images (Glannon, 2009, p. 325). Empirical concerns regarding neuroimaging experiments include experimental designs, ecological validity, individual variability in functional architecture, reverse inference reasoning, and the fallacy of localization—which we will return to later (Changeux et al., 1973; Derbyshire & Raja, 2011; Hardcastle, 2018, p. 187; Powell & Derbyshire, 2018, p. 358; Scarpazza et al., 2018; Tallis, 2016).

Further, the absence of a comprehensive model of the brain poses significant challenges in grounding claims, both conceptually and theoretically. Translating descriptions of neural events, signals, circuits, and correlates into higher-level concepts such as conscious awareness, cognitive processes, affective states, beliefs, intentions, rationality, and personal identity presents additional conceptual hurdles (Hardcastle, 2020, p. 157; Lewis, 1970, 1972; Lewis, 1966). These challenges are further complicated by considerations of epistemological accessibility and the enduring philosophical puzzles surrounding intentionality, phenomenological properties and first-person experience of consciousness (Chalmers, 1996a; Nagel, 1974). Considering all of these factors, it is reasonable to approach discussions regarding the existence of a 'moral' or 'criminal brain,' as well as claims about safety or efficacy, with a healthy degree of skepticism.

A second concern stemming from the 'neuroscientific turn' is the propagation of 'neuroessentialism'. This belief system asserts that neuroscience offers a comprehensive lens through which to understand human life, attributing the entirety of self and mind to the workings of the brain. This leads to a form of 'mindless neuroscience' encapsulated by the notion that for all intents and purposes, "we are our brains" (Cooter, 2014; Leefmann & Hildt, 2018, pp. 15-16; Reiner, 2011; Satel & Lilienfeld, 2013; Schultz, 2015; Vidal, 2009).

In metaphysical and philosophical discourse and debates about the 'mind-body problem,' these sentiments reflect a crude species of reductionist and physicalist conceptions of the interrelation between mental states and physical states (Chalmers, 1995, 1996a; Kim, 2007, 2018; Kim et al., 2012; Nagel, 1974; Ryle, 1949). I acknowledge it is important to reject dualistic perspectives that view the mind as a 'ghost in a machine,'[40] some mystical substance independent of the physical world. Instead, we begin from the proposition that we are constituted by our brains, but we cannot be solely defined in terms of them.

While the brain is an essential biological substrate with a privileged status in our mental life, it alone cannot fully account for the functioning of the mind, even at a physical level, nor

essential aspects of our mental life, our actions, behaviours, decisions, capacities for critical and moral reasoning, and rational agency—the latter a significant focus of discussion in this dissertation. At all junctures, it is crucial to acknowledge "It is not the brain but the subject constituted by the brain and the mind who is the agent" (Draper, 1974; Fuchs, 2004; see also Gallagher, 2005, p. 151; Glannon, 2009, p. 322). As I hope to show, understanding the mind in this way also better encapsulates a scientific view of what neuroscience tells us about how the brain operates in reality.

Setting these concerns aside, for contemporary neuroscience and what lies ahead, it is fair to characterize the state of scientific knowledge as 'nascent but promising' (Baskin et al., 2007, p. 239), with the recognition the upcoming 50 years are anticipated to significantly shape our understanding and application of neuroscience (Cara M. Altimus et al., 2020). I believe it is appropriate, with careful scrutiny, to view neuroscience as a discipline that provides tools that assist in part of a broader voyage of self-discovery, which can help shed light on "the kind of creatures we are" and provide insights into "the social structures that we inhabit and create" (Chiong, 2020; Farahany & Ramos, 2020, p. 152; Levy, 2007, p. 8; McCoy et al., 2020; Roskies, 2020a; Zawadzki & Adamczyk, 2021).

Beyond its clinical applications, neuroscience offers us an invaluable opportunity for critical self-reflection, allowing us to carefully question and challenge our conceptions of ourselves—our identity, nature, and the essence of who and perhaps even *what* we are. This process enables us to align our understanding with the insights provided by scientific inquiry. We have already begun to follow these themes in the previous chapter and will continue to do so in what follows.

Keeping this in mind, I now delve into an exploration of what current neuroscience reveals about the human brain, its functioning, and the valuable insights it can offer. This investigation sheds light on the subject at hand and enriches the discussion within this dissertation. And as we will see, when it comes to the brain, 'truth is stranger than fiction.'[41]

Cosmic Complexity: Unveiling the Human Brain

Neurointerventions distinguish themselves through direct operation on the brain, with our specific focus being the identification and 'treatment' of the biological basis of criminal activity. Critical questions thus arise: where within the vast complexities and active workings of the brain

can we pinpoint these features? What range of assumptions would we need to consider when claiming that a particular technology deployed would be safe and effective? To engage with these questions in a more meaningful sense, it is necessary to first scrutinize a set of fundamental facts about the brain.

The Hidden Symphony: the Enigmatic Components of the Brain

Encased in bone and nestled within the protective confines of the skull, the human brain, an extraordinary organic entity, resides as the seat of our consciousness. At its most basic level, 'The brain is a complex system of interconnected parts' (Siegel, 2012, p. 15). However, stating that the brain is complex is an understatement. Indeed, the brain is the most intricate structure in existence, whether of natural or artificial origin, in both our world and the known universe. Its complexity is such that, as Lilienfeld et al. (2018, p. 247) note, 'the person it inhabits has a difficult time fathoming the astonishing intricacy of its own architecture and functioning.'

The human brain, a compact entity weighing approximately three pounds, comprises roughly 86 billion neurons and trillions of glial cells.[42] Each neuron, on average, forms connections with ten thousand other neurons, amounting to an estimated one quadrillion (1,000,000,000,000,000) connections.[43]

The potential number of neuronal firing patterns at any given instant is a staggering ten to the power of one million—a number so large that it would require about 4000 pages to fully transcribe, with each page filled with zeros.[44] Across these myriad firing configurations, every second witnesses one hundred thousand chemical reactions.

Moreover, synaptic interconnections persistently adapt throughout the human organism's lifespan, making the range of patterns over a lifetime essentially limitless (Bear et al., 2020; Kandel et al., 2000; Koroshetz et al., 2020, p. 140; Lilienfeld et al., 2018, p. 247; Siegel, 2012, pp. 15-16).[45] We return to the issue of complexity later in this chapter.

Unifying Forces: Integration, Self-Regulation, and the Essence of Irreducible Properties

The principal function of the brain lies in receiving, processing, and directing the flow of energy and information.[46] The paramount inquiry revolves around the brain's remarkable ability as a biological system to shape meaning from an otherwise chaotic flow of information[47] in a manner that gives birth to indispensable facets of our mental existence. These facets include conscious experience, memory, intentions, beliefs, desires, and human rationality. The central

feature of the brain that makes this possible is *integration*, which enables higher-level function and gives rise to the fundamental properties of our mental life.

The dynamic flow of energy or information within specific networks or spatial regions of the brain is widely postulated to exhibit correlations with distinct spatial regions or networks that govern specialized processes—call these "subsystems" or "modules."

Integration is the process through which individual elements and properties of the brain's systems maintain their unique characteristics while also establishing interconnectedness. Consequently, the brain can be best described as "an interconnected and integrating system of subsystems" (Siegel, 2012, p. 19). This process allows distinct components of the system to retain their specific features while also becoming linked, enabling the integration of diverse modes of information processing into a cohesive whole. The outcome is an open, dynamic, flexible, and adaptive system that facilitates a cohesive representation of the world and facilitates self-regulation—a crucial aspect underlying the development of the brain throughout the lifespan of the human organism (Friston, 2010; Glannon, 2020, p. 90; Siegel, 2012, p. 9; Sporns, 2010; Tononi & Sporns, 2003).

Through integration, what arises is a complex system that manifests various irreducible properties, which are often discussed within the realm of complex system theory (Bassett & Sporns, 2017; Godfrey-Smith, 1998; Sporns, 2016). Emergence manifests through complex interactions between components, leading to phenomena at the system level that cannot be explained solely by component parts of the brain. These properties include nonlinearity and self-organization. Emergence is reflected in a vast array of high-level phenomena, such as consciousness and intentionality, that cannot be explained without accounting for specific properties of the system as a whole. Nonlinearity involves bidirectional causation observed in emergent self-organizing processes (Guastello et al., 2008; Kornfield, 2009; Siegel, 2012, p. 23). The brain, body, and environment are dynamically coupled through continual cycles of action and perception and mutually shape each other facilitated by re-entrant loops that enable flexible behaviour and adaptability to the environment (Glannon, 2020, p. 90; Siegel, 2012, p. 27). Self-organization signifies the capacity of the brain to autonomously align its components and intentional patterns without external guidance—a feature we focus on extensively in the next chapter. In essence, integration asserts that despite our ability to concentrate on individual brain regions, the entirety exceeds the aggregation of its components.

Exploring Functional Realms: Embodied, Enacted, Extended, and Embedded Cognition

What we will call the 'mind' is a part of this system, with distinct, irreducible properties; the mind is not reducible solely to the brain—the organ enclosed within the skull, encompassing the hierarchy of components, including single neurons, neuronal groups, circuits, systems, regions, hemispheres, and the entirety of the brain itself. Instead, mental states and human cognition operate across various functional domains—embodied, enacted, extended, and embedded (Ross et al., 2007; Varela et al., 2017).[48]

The mind is *embodied*, meaning its functions extend beyond the skull and are mediated through connections with other bodily systems—neural components within the skull intricately interact with neural, immune, endocrine, metabolic, cardiovascular, and musculoskeletal processes throughout the body.[49]

The mind is *enacted*. Brain states and mental states become enacted through bodily movements, directing actions and motivating movements by interacting with the physical environment. Through action, the environment influences the brain, and the brain, in turn, influences the environment.

The mind is also *extended*. It extends beyond the confines of the individual's body, intertwining with the physical environment. Mental processes transcend the boundaries of the brain and body, integrating external resources and artifacts into the cognitive system. These resources, such as tools or technologies, become indispensable for cognitive functioning, alleviating certain cognitive tasks (Clark & Chalmers, 1998; Damasio, 1994; Heersmink, 2016; Heinrichs, 2018, p. 60; Hurley, 1998).

Finally, the mind is *embedded*. This concept focuses on the idea that cognitive processes are embedded within a broader sociocultural and ecological context. It recognizes that cognitive processes are influenced by social interactions, cultural practices, and environmental factors—a concept sometimes referred to as 'anthropological embedment' (Kalis et al., 2008, p. 22; Rupert, 2009; Thompson, 2010; Varela et al., 2017).

Again, the relevance of functional domains and extended and embedded cognition are matters we focus on extensively for the balance of this dissertation in addressing conceptual and ethical issues, and we will return to them in what follows.

Bridging Minds: Unveiling the Social Fabric Woven Within the Brain

Incidentally, we must conceive of the brain as a "relational" and "social organ" (Fuchs, 2004, 2008, 2011; Fuchs & Schlimme, 2009; Glannon, 2009). As an integral part of a larger system, the brain processes neural signals from other brains, with relationships shaped by emergent, integrated, and non-reducible properties across enacted, embedded, and extended realms. Interactions with social and cultural environments influence the brain's structure, function, and development—and, in turn, are influenced by it. It is thus illogical to consider our relationships with others as separate from the system as a whole. These relationships mould the neural structures that enable the creation of a coherent worldview and the mental construction of reality (Glannon, 2018a, p. 330; Siegel, 2012, p. 9).

The significance of our connections with others has profound implications for brain development through a phenomenon known as "activity dependence" (Koroshetz et al., 2020, p. 140; Siegel, 2012 see generally Ch. 1). The brain's structure and function are constantly in a state of flux, shaped by social and environmental influences. These connections undergo changes through neuroplasticity, particularly during the early years of development, a process by which neurons and their connections are formed, lost, strengthened, or weakened based on experience (Fox et al., 2010; Kessler et al., 1997, p. 457). Our relationships with others play a pivotal role in structuring regulatory systems, thereby causing lasting impacts throughout our lives (D'Angiulli et al., 2008; Fox et al., 2010; Kessler et al., 1997, pp. 460-461; Kishiyama et al., 2009; Siegel, 2012, p. 22; Stevens et al., 2009).

We have seen issues about how the social and relational aspects of the mind are relevant in the last chapter, and we will further highlight their profound relevance in forthcoming discussions about how they might inform ethical issues surrounding brain intervention and the larger social and normative discourse.

Mental Time Travel: the Dynamic Dimensions of Cognition

Understanding how the brain mediates and amalgamates information across these realms requires taking time to dissect the temporal aspects of mental states. The central mechanism facilitating this process is memory—a 'dynamic process of continuous life' that allows individuals to formulate a coherent perception of the world and their place within it. This includes declarative memory,[50] which encompasses the recollection of particular events, factual

knowledge, and conceptual understanding. Memory possesses a retrospective aspect and a prospective dimension, essential for maintaining a personal sense of continuous existence through time (Glannon, 2019b, p. 5; Veselis, 2017).

The default mode network (DMN), a complex brain network, plays a significant role in the temporal aspects of cognition. It activates during periods of rest, facilitating a form of 'spontaneous internal mentation.' This leads to autobiographical remembering and envisioning of the future that enhances our understanding of others' thoughts and perspectives (Buckner et al., 2008, p. 20; Buckner & Carroll, 2007; Ingvar, 1979; Svoboda et al., 2006).

The combined effect of this is a unique feature of our mental life, the ability for 'mental time travel' located at a subjective time other than the present and accompanied by the first-person perspective (Michaelian, 2016; Michaelian et al., 2016; Zawadzki & Adamczyk, 2021, p. 8). We consider these features and temporality in greater depth in the next chapter, which discusses the parity principle and the scope of human rationality.

Exploring the Hierarchy of Brain Functions: From Primitive Origins to Advanced Cognition

By following the evolutionary paths encompassing two major brain areas, we gain insight into a hierarchy of brain functions and regions. This is based on the concept of 'vertical functional organization' or the 'cortico-subcortical hierarchy.' This hierarchy represents the organization of brain regions based on their evolutionary development and functional complexity (Fuster, 2001; Mesulam, 1998; Park & Friston, 2013).

The lower, 'subcortical regions' or deeper parts of the brain make up the 'ancestral mammalian brain.' They include primitive neural mechanisms for life sustenance, seen even in our earliest evolutionary predecessors. This area houses the hypothalamus and pituitary, crucial for maintaining physiological states like blood pressure and respiration, managing basic energy flow, and ensuring physiological balance, and governs survival responses and fosters the attachment system necessary for offspring protection (Herman et al., 2016; Striedter, 2005). This region is also often associated with 'affective' or 'emotional' responses.

The connections between cortical and subcortical regions are mediated through various systems. One primary area of research involves nerve fibres known as the cingulum bundle, which is a major pathway for communication between these two regions, helping to regulate

emotional responses and playing a role in cognitive functions such as attention and memory (Leech et al., 2011; Paul et al., 2019; Qadir et al., 2018).

Ascending the hierarchy from the 'ancestral brain' through the limbic system, we reach the upper cortical regions or the 'upper structure,' the distinctively human elements of our brain. Described as 'the seat of understanding,' these areas distinguish us from our primitive ancestors and are credited with empowering humans to create civilizations characterized by art, science, culture, and social institutions (Siegel, 2012, pp. 17-18; Tancredi, 2005, pp. 37-38). Included in these advanced areas is the cerebral cortex, or "neocortex," which contains functionally specialized areas possessing unique information ranges (Davies, 2020, p. 329; Dehaene & Changeux, 2004). The integration of these areas is essential for higher-level mental functioning, like consciousness, logical thought, and rationality, which form the basis of social and cooperative behaviour. Furthermore, the cortex, particularly the prefrontal cortex, orchestrates 'top-down control of behaviour' by balancing rational capacities with more primitive operations in the lower limbic areas (Damasio & Damasio, 2021; Dehaene, 2014; Glannon, 2019b; Goldberg, 2001; Miller & Cohen, 2001; Reghunath & Ghasi, 2020; Siegel, 2012, p. 18; Vogeley et al., 1999).[51]

It is vital to keep in mind that integration suggests psychological functions are spread across numerous brain networks rather than being confined to isolated regions (Dobbs, 2005, p. 24; Lilienfeld et al., 2018, p. 253; Weisberg et al., 2008).[52] However, this wide-ranging division serves as a useful heuristic for discussion in the next chapter, where we discuss how the structural connections between lower and higher cortical structures might inform us about the scope or limits of human rationality and the nature of cognitive and heuristic biases (Davies, 2009, 2020; Panksepp, 2004, 2012; Panksepp & Watt, 2011; Wilson, 2002).

The Global Workspace and the Adaptive Unconscious

The investigation of neural substrates of consciousness stands as a vital area of interest in cognitive and theoretical neuroscience, as well as in neuroethical discourse. The 'hard problem' of consciousness has historically fueled extensive debate in the philosophy of mind and precipitated a deep division in theoretical debate (Chalmers, 1996b, 2007; Churchland, 1989; Damasio & Damasio, 2021; Dennett, 1993; Kim, 2007; Nagel, 1974; Singer, 2019). This divide persists scientifically, with numerous contrasting theories evolving to elucidate the neural

correlates of consciousness (NCC) (Chalmers, 1996b; Dehaene, 2014; Facco et al., 2015; Kent & Wittmann, 2021; Laureys, 2015; Luby, 1998; Northoff & Lamme, 2020; Salles et al., 2018, p. 206; Tononi et al., 2016).[53]

To lay the groundwork for future chapters, we turn to the Global Workspace Theory (GWT) of consciousness. GWT posits a 'workspace' in the brain, spanning multiple regions—primarily in higher brain regions—that receives and 'broadcasts' information throughout various other regions (Mashour et al., 2020, p. 776). The GWT proposes that certain neural mechanisms within the workspace amplify signals until a self-sustaining 'ignition' occurs—aptly termed the 'combustion theory of consciousness' (Davies, 2020, p. 329; Dehaene, 2014; Dehaene & Changeux, 2004; Dehaene & Naccache, 2001). This ignition corresponds to what we subjectively perceive as a conscious state, often described as 'consciousness-as-reportability.' What is essential is that the 'signal' only rises to the level of conscious awareness when it is widely broadcasted across the workspace (Baars, 1993, 1997; Davies, 2020, p. 329; Dehaene, 2014; Dehaene & Changeux, 2004; Dehaene & Naccache, 2001; Mashour et al., 2020).

In the next chapter, we delve further into issues about consciousness which are crucial for neuroethical debates on how environmental factors can bypass our conscious awareness and the implications for parity reasoning, and compare the effects of conventional forms of punishment and the effect of novel neurointerventions (Bublitz, 2020b; Davies, 2009, 2020; Levy, 2020; Panksepp, 2004, 2012; Panksepp & Watt, 2011; Thaler, 2008; Wilson, 2002).

Exploring the Neurobiology of Morality and the Concept of the 'Criminal Brain'

Having set out some basic facts about the brain and mind, I now turn to issues about the 'moral' and 'criminal brain'—useful placeholder concepts, but as we have seen, perhaps ones that vastly oversimplify matters.

In 2005, renowned neuroscientist James Fallon used brain scans like PET and SPECT to study the brains of criminal offenders exhibiting psychopathic traits—including neuropathology and brain morphologies, which we will consider below (Muller et al., 2008; Nummenmaa et al., 2021; Tiihonen et al., 2008; Yang et al., 2009); (Decety et al., 2013; Deming & Koenigs, 2020, p. 1; Meffert et al., 2013). During his research, Fallon included his own brain scan as a

control. To his surprise, when comparing his scans to those of psychopathic individuals, he found similar brain patterns associated with psychopathy. This revelation prompted Fallon to explore his family history and personal traits further.

What is important is that, unlike the violent psychopathic individuals whose brains he had studied, Fallon had led a highly productive and successful life—which, through a process of self-discovery, he credits, in large part, to socioemotional factors, including a relatively stable upbringing (Fallon, 2014). This illustrates themes we have considered to this point—we are not our brains—and emphasizes the significance of investigating multiple aspects and domains when researching the neural foundations of morality and the complex relationship between brain mechanisms and criminal behaviour.

Prior to considering discussions concerning the neurobiology of morality, it is important to recognize the inherent complexity and elusive nature of the concept of 'morality' itself (Sinnott-Armstrong, 2008). While I consider various conceptions in what follows, here, I discuss it in a naturalistic sense as a collection of psychological adaptations that enable self-interested individuals to benefit from cooperation. It encompasses interwoven values, virtues, norms, practices, identities, institutions, technologies, and evolved neurobiological mechanisms. Together, these elements suppress or regulate self-interest, facilitating the formation of cooperative societies (Greene, 2013; Scheutz & Malle, 2018, p. 363; Sinnott-Armstrong, 2008).

Unravelling the Neurobiology of Morality—Ethical Dilemmas

Breakthroughs in neuroscience and neuroimaging have provided valuable insights into the complex neural processes underlying moral cognition, commonly referred to as the 'moral brain' (Moseley, 2020; Wheatley & Decety, 2015; Wiseman, 2016). A useful entry point for discussion is the iconic experiments conducted by Joshua Greene and colleagues that played a pivotal role in sparking contemporary discussions and developing theories regarding the neurobiological mechanisms involved in moral decision-making (Greene, 2003; Greene & Cohen, 2004; Greene, 2013; Greene & Haidt, 2002; Greene et al., 2004; Greene & Paxton, 2009; J. D. Greene et al., 2001).

These experiments used fMRI to track the brain activity of subjects while making moral and non-moral, personal and impersonal decisions modelled on the classic philosophical thought

experiments, the 'Trolly Problem' and 'Footbridge Dilemma' (Foot, 1967; Singer, 2005; Thompson, 1985; Unger, 1996).[54] Greene and his colleagues observed that distinct brain networks were activated depending on the nature of the moral dilemma and the specific moral judgments made by individuals.[55]

For example, when a subject responded using a 'utilitarian calculus'—pulling a lever to redirect a trolly to kill an overweight male to save five people—they observed a high degree of activation concentrated in the upper cortical regions of the brain, particularly the prefrontal cortex. Recall that this is the region of the brain often associated with 'higher-level' functioning and cognitive processing.

In a modified version, the 'Footbridge Dilemma,' subjects were asked to push the man off a footbridge and into the trolley's path, killing him but stopping the trolley. Most subjects experienced intense revulsion and could not act instead of allowing five other people to die. As they made this decision, parts of the limbic system were activated, most notably the amygdala; noted above, this region of the brain is associated with 'lower-level' emotional and affective processes. The researchers also observed that decisions involving the utilitarian calculus took a long time to process. But those involving emotional responses were much quicker (Greene, 2003; Greene & Cohen, 2004; Greene, 2013; Greene & Haidt, 2002; Greene et al., 2004; Greene & Paxton, 2009; J. D. Greene et al., 2001).

Building on these studies and predecessors, the 'Dual Process theory' has gained prominence in explaining moral cognition, supported by extensive research across multiple disciplines. The theory holds that the mind consists of two systems—System 1 and System 2—that play distinct but related roles in the cognitive processes underlying moral decision-making (Bublitz, 2020b, p. 64; Greene, 2013; Haidt, 2001; Kahneman, 2011).

System 1, referred to as the 'automatic setting,' encompasses a mode of 'fast thinking' that relies on neural processes that are evolutionarily primitive. It involves rapid, relatively emotional, and automatic mental processing, relying on heuristics and simple information processing strategies. It has been described as "nonanalytical, spontaneous, and impulsive" and is often adopted where "sensory inputs and environmental challenges require quick responses" (Banja, 2018, p. 287; Bublitz, 2020b, p. 64; Haidt, 2001; Kahneman, 2011); (Greene, 2013, p. 133).

In contrast, System 2, known as the 'manual mode,' engages in 'slow thinking.' It involves higher-level executive and intellectual functions, logical thought, and judgment. System 2 utilizes analytic and reflective thought processes, leading to more complex and diverse responses (Banja, 2018, p. 287; Bublitz, 2020b, p. 64; Haidt, 2001; Kahneman, 2011).

The Dual Process Theory of moral decision-making postulates that people enlist both System 1 and System 2 in resolving moral dilemmas depending on the circumstances in question. While both systems interact to integrate information, the "details of their interplay are unclear" (Evans & Frankish, 2009), as is the degree to which these systems map onto the divide between conscious and unconscious control (Bublitz, 2020b, p. 65; Frankish, 2009).

The Neurobiological Foundations of Human Morality

Neuroscience and neuroimaging breakthroughs have illuminated the complex neural processes of moral cognition, known as the 'moral brain.' Greene's experiments highlight specific brain regions, like the limbic system, while the System 1 and System 2 framework aids in categorizing cognitive processes.

Stemming for our discussion of cortical hierarchy, a useful starting point to understanding the neurobiological basis of morality is to consider the origins and history of the human species—which we touched on in the last chapter. The human capacities and social practices we call morality are mediated—but not reducible to—evolved neurobiological mechanisms that have played a distinct role in facilitating cooperative behaviours for survival and well-being throughout the history of our species (Siegel, 2012, pp. 17-18; Tancredi, 2005, pp. 37-38).

The origins of morality, in this understanding, can be traced back to pre-hominid mammalian species that possessed skills, practices, dispositions, and primitive attitudes promoting group survival long before Homo sapiens emerged in Eastern Africa. Alongside the emergence of human beings—recognizable as such—human societies were able to establish shared beliefs, principles, customs, and practices to uphold evaluative commitments and regulatory rules for their welfare (Banja, 2018, p. 288). These reflect many of the practices that today, we classify conceptually as 'morality.'

However, as mentioned in Chapter 1, humans possess neural mechanisms that align with a form of "common-sense" or "folk" morality. This refers to a set of moral attitudes that are commonly observed across diverse cultures and possess inherent capacities for promoting

cooperation among self-interested individuals (Churchland, 2011; Persson & Savulescu, 2008, 2011b; 2012, p. 10). As demonstrated, these reflections are also evident in responses to norm violations and the practice of punishment. This highlights concerns regarding how our primitive neurobiology contributes to a type of "moral tribalism" that may not be well-suited for addressing ethical issues in contemporary human societies (Greene, 2013; Persson & Savulescu, 2012).

Isolating the 'Moral Brain'

To identify the 'moral brain' and examine the prospects of intervening in it, it is crucial to revisit the aforementioned principles. Moral reasoning is an outcome of a dynamic, extended, and integrated system that possesses irreducible properties. It is most aptly regarded as an emergent self-organizing process that, within this comprehensive system, is tied to specific subsystems, including a variety of cognitive and affective processes, realized through action and situated in a larger social and normative framework.

I argue it follows that corresponding brain-based mechanisms are just one component of this system. In reality, the key factors underlying morality in the brain are the mechanisms that enable integration and facilitate self-regulation and irreducible properties. Particularly important are the dynamic features that influence social and relational aspects, especially during early development, where neural pathways are formed across different brain regions to establish specialized networks responsible for fostering healthy connections and social behaviours. As I described in the last chapter, these neural substrates and integrated systems of subsystems play a crucial role in forging the 'bonds that unite humankind.'

Notwithstanding, there are various brain-based areas of inquiry that have captured attention in contemporary neuroscience—although, building on our previous discussions, I will question the utility of a prevailing paradigm that places undue weight on these features.

The first involves brain networks associated with cognitive functions that facilitate rationality, regulation, and impulse control. Building on the general principles of 'cortico-subcortical hierarchy' above, at the most general level, many of these mechanisms facilitate 'top-down' regulation from higher brain regions—tracing themes of System 1 processing in Dual Network theory. In technical terms, this often centres discussions on a brain region known as the Prefrontal Cortex (PFC), known to play a crucial role in higher cognitive functions, such as

problem-solving, planning, and decision-making (Joshua D Greene et al., 2001; Koenigs et al., 2007).

The Posterior Cingulate Cortex (PCC) has been linked to autobiographical memory retrieval and envisioning the future, both of which can contribute to moral decision-making. For example, recalling past moral or immoral actions may influence present moral judgments (Buckholtz & Marois, 2012; Harenski & Hamann, 2006).

Another region is a subset, the Ventromedial prefrontal cortex (vmPFC), implicated in social and moral cognition. This region is thought to be involved in integrating multiple types of information and anticipating the consequences of one's actions. (Greene et al., 2004; Moll et al., 2003).

These higher cognitive functions interact with neural mechanisms located in lower cortical regions that regulate emotion, often classified under the ambit of 'affective processes'—a broad range of emotional and motivational functions involved in experiencing and regulating emotions. (Blair, 2007; Harenski & Hamann, 2006).

The integration of these various features plays a significant role in moral reasoning, particularly in the development of empathy and 'theory of mind,' which are closely connected concepts in the realm of social cognition. Empathy refers to the ability to understand and share the feelings of others, while theory of mind refers to the capacity to attribute mental states, beliefs, and intentions to oneself and others.

These concepts are interconnected because theory of mind enables individuals to understand and predict the mental and emotional states of others, which forms the basis for empathetic responses. Both utilize higher-level executive processes to integrate knowledge about others' thoughts with information about consequences and emotions during moral judgment, reflecting "one of the most fundamental, and arguably uniquely human, aspects of social and cognitive development" (Decety & Howard, 2013, p. 53; Powell & Derbyshire, 2018, pp. 353-354; Wellman, 2011).

Again, it is important to remember that these brain regions do not work in isolation and form part of a complex network where there are likely to be other regions and mechanisms involved. For example, the very basis of 'theory of mind' and the possibility of 'empathy' involves an understanding of social aspects of the mind, in addition to embedded considerations inherent in moral and normative judgements about 'appropriate' moral responses. Even if each

region were accounted for, it still would leave out the dynamic interrelation between the enviro-societal community within which the brain persists.

The incredible complexity of the brain and its relationship to the mind, within the multifaceted dynamics of integration, does not allow for a simple linear nor fixed categorization of the moral workings of the brain. We are only beginning to understand how these intricate neural processes contribute to our moral judgements and how they can go wrong.

The Criminal Brain

This dissertation investigates the moral acceptability of using advanced technologies to intervene in the brains of criminal offenders—the 'criminal brain.' In addition to studying the neurobiology of morality, research has identified neural substrates associated with antisocial behaviour and violent aggression—that contribute to criminal behaviour. These findings can provide insights into underlying dysfunctions that aid in identifying, explaining, or predicting the occurrence of crime. Again, a theme we will return to is even if we accept broad placeholders, the 'moral mind' and what we see as the 'criminal mind,' the former is premised on particular moral and normative judgements, and the latter, a social or political classification about human behaviour.

A useful starting point is psychopathy, a personality disorder marked by enduring antisocial behaviour, aggression, remorselessness, and, notably, an empathy deficit, which serves as a prime example. Psychopaths can cognitively grasp others' emotions but fail to resonate with their distress or exhibit empathic concern (Hare, 1991, 2003). This disorder significantly predicts criminal behaviour, violence, and the likelihood of violent reoffending, with approximately a quarter of adult male prisoners being psychopaths, though they only constitute 1% of the male population (Deming & Koenigs, 2020, p. 1; Hare, 2003). Notoriously resistant to treatment, psychopathy exhibits structural and functional brain anomalies, including diminished amygdala activity and deformations in the PFC, vmPFC, and limbic regions (Aylett et al., 2006; Blair et al., 1996; Cooke & Michie, 2001; Deming & Koenigs, 2020; Hare, 2003; Harris et al., 1991; Kiehl & Hoffman, 2011; Muller et al., 2008; Nummenmaa et al., 2021; Poeppl et al., 2019; Powell & Derbyshire, 2018, p. 355; Pujol et al., 2012; Tiihonen et al., 2008; Yang et al., 2009).

Interestingly, though psychopathy often has a genetic component, it's not solely dictated by our DNA. Epigenetic factors can play a critical role, particularly in the expression of high-risk genes such as Monoamine Oxidase A (MAOA). This includes factors such as stress or abuse early in life. These factors can steer whether genetic predispositions translate into violent behaviour (Deming & Koenigs, 2020; Gao et al., 2009; Kiehl, 2006)—recall the case of Fallon.

Along these lines, aside from prototypical cases like psychopathy, other brain malfunctions contribute to criminal offending, particularly disruptions during critical developmental stages. Childhood, an essential period for developing and strengthening neurobiological connections, is susceptible to adverse conditions such as poor nutrition, physical abuse, violence exposure, neglect, and traumatic life events, all of which can lead to emotional regulation and social behaviour problems. This includes the failure to develop 'theory of mind,' and an absence of empathy, leading to "callous, criminal, and antisocial behaviour" (Aylett et al., 2006; Blair et al., 1996; De Gelder et al., 2004; Deming & Koenigs, 2020; Harris, 2003; Hess & Blairy, 2001; Hillis, 2014; Poeppl et al., 2019; Powell & Derbyshire, 2018, p. 357; Richell et al., 2003; Shamay-Tsoory et al., 2009; Widom, 1978).

These conditions are closely linked to reduced function and neuroanatomical abnormalities in particular regions of the brain. For example, hippocampus impairment, amygdala dysregulation, and dysfunctions in the Hypothalamic-Pituitary-Adrenal (HPA) Axis. These disruptions can generate various psychological disorders, including PTSD, depression, anxiety disorders, dissociative disorders, personality disorders, and substance use disorders (SUDs), which cumulatively increase the risk of criminal offending and reoffending (Bremner, 2022; Coppola, 2018, p. 5; Cozolino, 2014; D'Angiulli et al., 2008; Fox et al., 2010; Gillespie et al., 2017; Hart & Rubia, 2012; Kessler et al., 1997, pp. 460-461; Kishiyama et al., 2009; Koenigs et al., 2007; Kolk & Fisler, 1994; Leutgeb et al., 2016; Ma et al., 2011; McLaughlin & Lambert, 2017; Ostovar, 2009; Piotrowska et al., 2015; Raine, 2008; Siegel, 2012, p. 22; Sobhani & Bechara, 2011; Stevens et al., 2009; Teicher et al., 2012; Weller et al., 2007).

Cognitive defects, particularly in the PFC, PCC, and vmPFC—associated with executive functions—are risk factors for criminal behaviour due to their 'top-down' control over affective processes. As Glannon explains, "[t]he ability to regulate one's impulses depends on connectivity between the prefrontal cortex, which underpins rationality, and the limbic system, which underpins basic desires and emotions" (Glannon, 2020, p. 98). For example, clinical disorders

such as ADHD are linked to neurotransmitter dysregulation and smaller overall brain volume in areas related to attention and impulse control, such as the PFC and cerebellum (Cortese et al., 2012; Hoogman et al., 2017) and the prevalence of ADHD among male inmates is between 25% and 40% (Ginsberg & Lindefors, 2012; Ryberg, 2020, p. 7).

The same holds for disorders such as SUDS, a significant factor in incarceration, involve dysfunctions in limbic system areas crucial for 'top-down' control (Clark et al., 2004; Davis et al., 2013; Khantzian, 1997; Regier et al., 1990). Similarly, an area of interest is disruptions, including reductions in white matter integrity, in regions such as the cingulum bundle and cingulate gyrus, that connect higher and lower cortical functions, which have been implicated in a variety of mental health conditions, including depression, anxiety, and post-traumatic stress disorder, and in conjunction with other structural and functional abnormalities, contribute to impulsive-antisocial psychopathic traits (Leech et al., 2011; Paul et al., 2019; Qadir et al., 2018). Again, this ties in with themes we explored in the previous Chapter. But it also highlights environmental factors play a vital role in understanding antisocial or criminal behaviours.

Intervening in the Criminal Brain—Emerging Technologies

In exploring the intricate links between the brain, morality, and criminal behaviour, we are witnessing a transformative era with the rise of groundbreaking technologies that aim to manipulate these processes. This section introduces and examines novel neurointerventions, highlighting their enhanced effectiveness and potential application within the criminal justice system. While technical details are not extensively covered, the subsequent discussion provides relevant information on the complexity of these interventions, addresses empirical and conceptual issues, and serves as a foundation for ethical deliberations—particularly those related to the promise and perils they pose.

This dissertation focuses on what I have defined as 'neurointerventions'—as they are frequently but not exclusively referred to in the literature surrounding punishment (Birks, 2018; Douglas, 2014b; Holmen, 2020; Matravers, 2018; Nadelhoffer et al.; Ryberg, 2020, 2021; Shaw, 2018; Vallentyne, 2018a; 2018b, p. 124; Nicole A Vincent et al., 2020b). I adopt a broad definition of neurointerventions as "chemical, electrical, surgical, and other interventions that act directly on the brain" (Vallentyne, 2018b, p. 124). This involves a crucial distinction

between the 'brain'—the physical organ contained within the human skull—and the 'mind,' with the former being integral to, yet not synonymous with, the latter.

The rise of these technologies has sparked a contentious practical and ethical discourse, not just in the criminal context but across various other domains, including the 'enhancement debate,' which revolves around the utilization of such technologies to enhance brain function rather than solely treating brain dysfunction (Carman, 2021; Earp et al., 2018; Harris, 2011, 2014a; Persson & Savulescu, 2008, 2011a, 2011b, 2012, 2013, 2015; Savulescu & Persson, 2012; Sparrow, 2013; R. Sparrow, 2014; R. J. Sparrow, 2014; Wiseman, 2016) (Buchanan, 2011; Focquaert & Schermer, 2015, p. 139). We return to this in the chapters that follow, as it highlights many of the issues considered here.[56]

At the outset, it is important to note that there is nothing inherently new about 'neurointerventions' as defined here. As Merkel explains, Ancient Greek writings by Galen and Dioscorides mention the administration of opium for sleep disorders and pain relief (Merkel et al., 2007, p. 11). In prehistoric times, substances like opium, cannabis, peyote, and alcohol were used in cultural practices, including those involving hallucinogenic effects (Earp et al., 2018, p. 175; Homan, 2011; McKenna et al., 1984).[57] Neanderthals may have consumed plants with amphetamine-like effects as early as 50,000 BCE (Merlin, 2003; Wolpe, 2018, p. 220). Controversial theories even suggest that early hominids' consumption of psilocybin-containing fungi played a role in our adaptive evolutionary history, including brain evolution (Rodriguez Arce & Winkelman, 2021).

In the present day, various studies have shown that substances such as caffeine and glucose impact cognitive function (Conan, 2020; Kitajka et al., 2004; Yehuda et al., 2005). while vitamin supplements and nutritional interventions have been associated with potential changes in brain structure and function (Gesch et al., 2002; Schoenthaler Stephen Amos Walter Do, 2009; Zaalberg et al., 2010). Omega-3 supplementation has also garnered attention for its role in supporting healthy brain function and memory (Kitajka et al., 2004), with studies indicating improved functioning in specific brain regions such as the dorsolateral prefrontal cortex (McNamara & Carlson, 2006). But while similar, in principle, it does seem there is something about contemporary neurointerventions that is distinct and sets them apart. As we will see, these interventions are uniquely characterized by their precise targeting, advanced

integration with real-time monitoring technologies, and stringent ethical standards, marking a new phase in the application of neuroscience.

Psychopharmacological Interventions

To aid discussions, novel technologies can be categorized into drugs and devices,[58] with each classification assessed separately in relation to their potential for preventing crime and reducing the risk of recidivism—as opposed to an extensive focus on clinical issues.

The use of drugs to modulate neurotransmitters can be a valuable consideration in psychopharmacological interventions. Various agents, such as antipsychotics, beta-blockers, testosterone, Levodopa, and Ecstasy, have been suggested to alter specific neurobiological mechanisms associated with dispositions like impulsivity, aggression, empathy, generosity, and cooperation (Berman et al., 2009; Crockett et al., 2010; De Deyn & Buitelaar, 2006; Pappadopulos et al., 2006; Pedroni et al., 2014; Terbeck et al., 2014; Turner, 2016) (Crockett & Rini, 2015; Dubljevic & Racine, 2017, p. 344; Hysek et al., 2014). Selective Serotonin Reuptake Inhibitors (SSRIs) have gained prominence in treating depression and anxiety by increasing serotonin availability. For example, a recent meta-analysis involving over 6500 participants and 175 studies found a small but reliable correlation between serotonin levels and aggression (Chew et al., 2018, p. 23).

One area of inquiry is the use of so-called ‘smart drugs’ to enhance cognition. Discussions traditionally focus on synthetic stimulants like amphetamine, piracetam, methylphenidate, and modafinil, which are used to treat psychopathologies such as attention deficit disorder and narcolepsy.[59] Some research suggests that these drugs correlate with improved cognitive capacities, especially in memory, attention, and executive functions, with minimal side effects (Elfferich, 2021; Moskal et al., 2014; Turner et al., 2003). Considering the prevalence of disorders like ADHD in prisons and the limitations in higher cognitive processes that can hinder self-regulation, cognitive enhancement is a strong candidate for application in criminal justice practices (Bülow, 2020; Ginsberg & Lindefors, 2012; Ryberg, 2020, p. 7).

Psychopharmaceuticals, including topiramate, have also gathered much attention. These developments have been used to treat substance use disorders (SUD) and are utilized in the therapeutic jurisprudence movement. Topiramate, used in drug courts, reduces substance

cravings and improves addiction treatment outcomes. Opioid agonists and methadone maintenance therapy (MMT) are also effective for treating SUD, specifically for drugs like methamphetamine and heroin. Extensive research, including observational studies and randomized control trials, supports the use of MMT in reducing drug relapse and improving treatment retention (Bahr et al., 2012; Chandler et al., 2009; Chew et al., 2018, p. 19; Dolan et al., 2005; Egli et al., 2009; Gordon et al., 2012; Hall et al., 1993; Hedrich et al., 2012; Sifferd, 2020, p. 309).

A final avenue of interest involves Antiandrogen treatment, also known as anti-libidinal pharmacological agents (ALPAs), which is administered to certain offenders to treat sexual paraphilias. These drugs are designed to lower testosterone levels or mitigate the effects of testosterone, commonly referred to as "chemical castration."[60] Cyproterone acetate (CPA) and medroxyprogesterone acetate (MPA) are prominent antiandrogen drugs used for this purpose and have been employed in certain jurisdictions, including Europe, Canada, and the US, as a means of crime prevention (Chew et al., 2018, pp. 13-14; Ryan, 2020; Nicole A Vincent et al., 2020a, p. 23).

Devices for Neuromodulation

Neurointerventions can also involve the use of neurodevices, which are devices that directly interact with the brain through electronic or magnetic pulses. These devices aim to modulate brain activity by either stimulating or inhibiting it, a process known as neuromodulation. Neuromodulation, involving diverse specialties, is a rapidly growing field impacting numerous patients worldwide, and technologies continue to advance rapidly and are widely considered "among the most powerful means currently available for intervening on the human brain" (Christen & Müller, 2017; Zuk et al., 2018, p. 49).

In 1938, psychiatrist Ugo Cerletti introduced a procedure called "shock therapy," involving the administration of transcranial electric shocks to induce seizures (Rzesnitzek & Lang, 2016). This procedure, also known as Electroconvulsive Therapy (ECT), is still used today under general anesthesia as a treatment for drug-resistant depression and related disorders. However, it remains controversial due to its historical association with its use in Nazi Germany as a form of euthanasia murder (Friedlander, 1995; Rzesnitzek & Lang, 2017).

However, this early example of neuromodulation sharply contrasts with sophisticated contemporary iterations. Two notable neurodevices are transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS), which have been extensively discussed in the neuroethics literature.[61] Both operate by altering cortical excitability or increasing the flow of energy and information in the brain and can target more precise brain regions instead of inducing an indiscriminate seizure (Bennabi et al., 2014; Liu et al., 2021).[62] While the efficacy of tDCS remains a topic of debate, there is a relative consensus it has "a minimal positive impact on the executive function" (Chhatbar & Feng, 2015; Coffman et al., 2014; Conan, 2020; Dedoncker et al., 2016; Hill et al., 2016).

Very preliminary research indicates that tDCS may have implications for managing addictions and reducing aggression, with studies suggesting reductions in self-reported aggression, changes in intention to commit assault, and increased altruism and trustworthiness (O. Choy et al., 2018; Conti & Nakamura-Palacios, 2014; Jansen et al., 2013; Molero-Chamizo et al., 2019). Moreover, tDCS applied to the prefrontal cortex has been investigated in relation to aggression, impulsivity, and moral judgment (Brevet-Aeby et al., 2016; Dambacher et al., 2015; Jeurissen et al., 2014; Molero-Chamizo et al., 2019; Riva et al., 2015; Tassy et al., 2012; Young et al., 2010).

Deep Brain Stimulation (DBS) is a neuromodulation technique involving the surgical implantation of electrodes directly into the brain. It offers greater precision and effectiveness in modulating activity in deeper brain regions compared to non-invasive techniques like TMS and tDCS (Glannon, 2014). DBS has shown promising results in treating Parkinson's disease, essential tremors, and epilepsy and is being explored for other disorders such as major depressive disorder, depression and Alzheimer's disease (Goering et al., 2021, p. 2; Maslen et al., 2018; Park et al., 2017; Siegel et al., 2017) (Beeker et al., 2017). Certain preliminary research suggests that DBS may impact regions associated with moral reasoning and judgment, aggression reduction, and managing sexual urges in paraphilic individuals (Franzini et al., 2013; Fumagalli et al., 2011; Fumagalli et al., 2015; Fuss et al., 2015).[63] Hypothetically, precise targeting of deeper brain structures, such as the limbic system—central to emotional regulation—could modulate emotional responses.

A final advancement of interest is optogenetics, a revolutionary technique that involves the genetic modification of neurons, making them responsive to light via the introduction of opsin

genes. This allows precise and *bidirectional* modulation of neuronal activity, a unique advantage over traditional neurostimulation methods. It can 'turn on' or 'turn off' cells in specific regions (Adamczyk & Zawadzki, 2020; Boyden et al., 2005; Deisseroth, 2011; Packer et al., 2013; Yizhar et al., 2011; Zawadzki & Adamczyk, 2021). Paired with neuroimaging technologies, it births a new method known as optogenetic functional magnetic resonance imaging (Lin et al., 2016; Mahmoudi et al., 2017).

Animal studies have provided exciting results, including the restoration of hearing in mice and memory retrieval in rats (Adamczyk & Zawadzki, 2020, pp. 209-210; Guskjolen et al., 2018). Intriguingly, stimulating certain brain regions can cause drastic behaviour changes, as seen in male mice that exhibited aggression when their ventromedial hypothalamus was activated, leading to attacks on female rats and inanimate objects. Notably, inhibiting these regions suppressed such aggression (Lin et al., 2011). With the first human trials approved, the full extent of optogenetics' implications remains to be discovered.

A final, more hypothetical realm of discussion involves the use of neural prosthetics and brain-computer interfaces (BCIs)—which could be incorporated with both novel brain imaging and other forms of modulation. Neural prosthetics employ electronic devices that are surgically implanted within the brain to compensate for impaired motor, sensory, or cognitive functions. These devices serve as functional substitutes for the damaged areas, thereby reinstating lost abilities or modulating aberrant activity (Burke et al., 2014; Glannon, 2020).[64] Neural implants aimed at the hippocampus and entorhinal cortex show promise for improving memory functions such as information storage and recall and are projected to be patient-ready in the near future (Berger et al., 2011; Glannon, 2020, p. 102; Hampson et al., 2013). While firmly rooted in science fiction, the concept of using brain implants and BCIs for interventions that could predict and prevent criminal acts, such as violence, has provided intriguing material for thought experiments (DeGrazia, 2014, p. 63; Ryberg, 2020, p. 63).

The purpose of introducing these technologies is to identify the current state of the science and how it informs issues about assumptions of safety and efficacy and, more generally, the sorts of issues we might face in the future. In many senses, the scope of hypothetical discussions has now expanded beyond the realm of science fiction. Neuroimaging technologies have advanced at a rapid pace, enabling increased precision in neuromodulation. Furthermore, the reciprocal and amplifying impact of these advancements, combined with the power of

artificial intelligence (AI), has revolutionized our understanding of the human brain and its potential for intervention. In tandem, our ethical theorizing must keep stride, highlighting the importance of imagination and the merits of ideal theorizing. However, before delving into these matters, it is essential to distinguish between the realms of possibility and imagination.

Assessing Contemporary Neurointerventions—Empirical Considerations

The goal of exploring the workings of the brain and the nature of novel technologies has been to ground theoretical discussions moving forward in a scientific understanding and the realm of practical possibilities. Birx and Buyx note that the majority of ethical debates tend to "idealize the effects of the neurointervention" (Birks & Buyx, 2018, p. 135). We have discussed various new neurointerventions that span a range from approved and operational to entirely speculative.

Punishment equivalence arguments aim to draw comparisons between ideal technologies that are 'safe and effective' and other conventional means used for punishment. However, the more we veer towards speculative technologies, the greater the number of assumptions needed, thereby widening the chasm for discussing feasible applications. I begin by addressing the safety and efficacy of current technologies before identifying what I see as significant barriers moving forward and certain misconceptions that underlie strong claims about the potential efficacy and future applications of these technologies.

Safety and Efficacy

For any applied discussion about immediate application, concerns about safety are paramount. I have adopted views of the brain as part of a larger integrated and extended system that gives rise to emergent properties, in large part, to enable flexible behaviour and adaptability to the physical and social environment. Herein, I think, lies many of the problems we see with contemporary interventions and technologies.

In medicine, it is widely recognized that medical interventions often lead to unintended and unforeseen side effects (Golan & Tashjia, 2004; Kline, 1959). With respect to the brain, complex and vital organ, "new psychoactive medications and the direct stimulation of the brain may have unforeseen and terrible consequences" (Elfferich, 2021, p. 132; Greely et al., 2008).

The effects and potential side effects of any form of treatment can vary significantly due to differences in functional architecture and individual variability.

For instance, although SSRIs have a small correlation to reduced aggression (Chew et al., 2018, p. 23), understanding their impact involves assessing the nuanced interplay of over 100 neurotransmitters, individual traits, environment, and social interactions (see generally Meriney, 2019). Further, chronic use of such drugs can lead to adjustments favouring the treated state based on intrinsic self-regulatory mechanisms, which favour synaptic homeostasis (Chew et al., 2018, p. 23).

Psychopharmacological interventions, such as topiramate and MMT, have certain side effects, such as weight gain, sleep disturbances, changes in sexual function, and withdrawal cravings. However, most significantly, clinical outcomes are heavily dependent on socio-emotional factors, such as voluntary engagement with cognitive-behavioural therapy (Bell & Strang, 2020).

Methylphenidate and other cognitive enhancers can improve cognitive functions, but actual improvements are seen to be far lower than assumed in theoretical discussions. They are more effective for individuals with lower baseline performance but can impair performance in healthy individuals and hinder cognitive flexibility (Dresler et al., 2019, p. 1139; Fallon et al., 2017; Finke et al., 2010; Schleim & Quednow, 2018).

Antiandrogen treatment, also known as 'chemical castration,' is associated with significant adverse effects. These include rapid bone loss, osteoporosis, cardiovascular problems, diabetes mellitus, fatigue, and enlargement of breast tissue.[65] But most significantly, their efficacy in reducing recidivism among sex offenders is a subject of great debate (Douglas et al., 2013; Kutcher, 2010b; Liberto, 2018; McMillan, 2014; Ryan, 2020; Scott & Holmberg, 2003; Sifferd, 2020). According to Christopher Ryan, based on a systematic review of the literature, it cannot be asserted that antiandrogen therapy has proven efficacy in reducing the rate of recidivism among sex offenders: "In all probability, the damn things don't work" (Ryan, 2020, p. 287). This points to the potential for strong emotional responses and societal pressures to shape policy decisions, perhaps at the expense of evidence-based practices—a matter discussed in Chapter 1.

Furthermore, insights from a qualitative study involving interviews with men convicted of sexual offences about their perceptions of quasi-coercive offers of biological treatment reveal

mixed outcomes (Knack et al., 2020). Many participants noted that while treatments like antiandrogen drugs helped suppress distracting thoughts, they were not adequate on their own to prevent recidivism. This reflects the complexities of determining the effectiveness of such treatments, as detecting all instances of recidivism is challenging, highlighting the limitations of current research methodologies. Participants also expressed concerns about the coercive aspects of these treatments when consent is legally incentivized, underscoring the ethical dimensions of using such interventions within the criminal justice system (Knack et al., 2020). This emphasizes the necessity of a more holistic approach that combines pharmacological treatments with psychological and social support to more comprehensively address the multifaceted nature of criminal behaviour.[66]

In the context of novel neuromodulation techniques, even precise application forms can affect brain tissue beyond the intended area due to the intricate interconnections of the brain's functional networks (Birks & Buyx, 2018, p. 135). While the safety of transcranial direct current stimulation (tDCS) has been demonstrated in controlled laboratory settings, long-term effects and real-world implementation require further study due to limited data on sustained treatment and side effects, and initial research suggests trade-offs with cognitive enhancements, where improving one function may come at the expense of others (Birks & Buyx, 2018, p. 135; Dresler et al., 2019, p. 1140; Dubljevic et al., 2014). Moreover, a recent systematic review underscores the necessity for cautious application of these techniques in criminal justice settings, revealing that while there are potential benefits, the methodological limitations of current studies and the complexity of neuromodulatory effects on behaviour warrant a more nuanced approach (Romero-Maronez et al., 2020). Another issue relates to the fact TMS and tDCS are ineffective in targeting many deep brain regions that are essential for cognitive and emotional processes (Thair et al., 2017).

More invasive surgical procedures like deep brain stimulation (DBS) and neural prosthetics have the potential to reach these deeper brain regions, they are highly invasive procedures and carry significant side effects on personality, behaviour, and impulse control, raising ethical concerns (Birks & Buyx, 2018, p. 135; de Haan et al., 2015; Gilbert & Lancelot, 2021, p. 20; Goering et al., 2021, p. 2; Klein et al., 2016). Similarly, cutting-edge technologies, including optogenetics, also involve a surgical procedure, and it has been cautioned use to alter

moral reactions and evaluative schemes may pose significant unforeseen implications (Adamczyk & Zawadzki, 2020, p. 208).

However, despite these challenges, recent advancements in non-invasive brain stimulation techniques, such as Transcranial Ultrasound Stimulation (TUS) and Temporal Interference (TI) electrical stimulation, offer promising alternatives. These methods have demonstrated the ability to effectively target and modulate deep brain regions like the limbic system without surgical intervention. TUS, for instance, achieves high spatial resolution deep within the brain (Darmani et al., 2022), while TI has been shown to enhance cognitive functions by focusing on areas like the hippocampus with minimal impact on surrounding tissues (Grossman et al., 2017; Violante et al., 2023). These emerging technologies not only reflect significant scientific progress but also address the ethical concerns associated with more invasive procedures, marking a pivotal shift in how deep brain areas might be accessed and treated in the future.[67]

From Practical to Ideal Theorizing

In discussing existing technologies, it is important to identify that ethical theorizing about specific technologies falls upon a spectrum which ranges from practical application to the more hypothetical and theoretical.

On the lower end of the spectrum of theorizing, there are practical discussions about the applications of existing technologies, for example, nutritional interventions such as omega-3 (Conan, 2020, p. 43) and psychopharmacological interventions, such as methylphenidate (Bülow, 2020),[68] some even suggesting tDCS has reached a point where it is practical and ethically feasible to seriously consider administering it in the prison context (Conan, 2020, p. 43).

Transitioning towards more speculative realms, closed-loop interventions represent a significant development in neurotechnology (Goering et al., 2017; Kellmeyer et al., 2017; Klein et al., 2016; Parastarfeizabadi & Kouzani, 2017). This advanced approach integrates continuous monitoring with responsive stimulation based on detected biomarkers or behaviours, potentially akin to speculative applications like modulating behaviours in individuals with gambling disorders when they approach a casino. This compares to systems like alcohol ignition

interlock devices that control behaviour through continuous surveillance,[69] yet it introduces a higher level of complexity and ethical consideration due to its direct intervention in neural processes.

As discussions become increasingly theoretical, examples gravitate towards cutting-edge applications such as wireless brain-computer interfaces utilizing artificial intelligence, and neurodevices designed to predict aggressive, violent outbursts (DeGrazia, 2014, p. 63; Ryberg, 2020, p. 12; Stefano, 2021, p. 213). These technologies include Brain-Computer Interfaces (BCIs) that forge connections between the brain and computers, significantly enhanced by the integration of biomedical technologies with artificial intelligence (AI) (Burke et al., 2014; Glannon, 2020, p. 102; Lebedev, 2014; Lebedev & Nicolelis, 2006; Steinert et al., 2018; Wolpaw & Wolpaw, 2012).[70] For instance, researchers have used EEG recordings and AI techniques to 'learn about the subject' and forecast and estimate the likelihood of seizures days in advance (Gardner et al., 2006; Gilbert & Dodds, 2020, p. 113; Proix et al., 2021).

Further, it is believed, at least 'in concept,' that such technologies could facilitate 'brain-to-brain interfaces' (BTBIs), a remarkable group of projects regarded as "among the latest, most impressive, and morally controversial" (Grau et al., 2014; Jiang et al., 2019; Li & Zhang, 2016; Li & Zhang, 2017; Pais-Vieira et al., 2013). Early studies have shown rats making behavioural choices influenced by BTBIs (Yoo et al., 2013), humans controlling rat tail movements via BTBIs (Yoo et al., 2013),[71] and successful guidance of 'cyborg cockroaches' around an S-shape track (Friedrich et al., 2018, p. 17; Li & Zhang, 2017).[72]

Notably, BTBIs have even facilitated basic information transfer between human brains, enabling simple actions using brain imaging and neuromodulation techniques (Pais-Vieira et al., 2013). Lastly, at the most extreme end, we will later explore a thought experiment involving an omniscient 'God Machine' that tracks people's thoughts and intervenes to prevent severely immoral actions (Savulescu & Persson, 2012, pp. 114-115).[73]

As we delve into the realm of present technologies and explore the topics discussed here, it becomes evident that the landscape of contemporary science is in a constant state of flux. The captivating aspect lies in the trajectory that spans from current possibilities to the far-reaching spectre of a dystopian future. This ever-evolving journey traverses the boundaries of scientific advancement, captivating our attention and fueling our curiosity. Engaging discussions emerge

around actual technologies and the evolving boundaries they present—and so too the range of corresponding ethical issues we must address.

Immensity of Complexity, Localization and the Boundaries of Human Comprehension

Building on these advancements, in addressing empirical issues about 'safety and efficacy and corresponding conceptual and ethical issues, I think it is safe to consider where we *might* go—and what obstacles could foreseeably prevent us from getting there. This falls in line with calls by ethicists that due to the evolving state of the science, neuroscience should be seen as a "predominantly anticipatory field" (Farah, 2011b, p. 776) aimed at considering "emerging applications of neurotechnology through a lens that seeks to identify and prepare for both benefits and potentially unwanted outcomes" (Farahany & Ramos, 2020, p. 150). Beyond available technologies, it is vital to address the limits of what we know and perhaps what we could even know in either the near or distant future: "Given the vast complexity of the human brain, what we can achieve in the foreseeable future is extremely limited" (Murphy & Greely, 2011).

To frame issues and consider the target of discussion—the brain—let us begin with the astronomical figures above and how they relate to our current technologies. Studies in the cognitive sciences suggest, by design, the very subject of study, the human brain and supervenient mind, is poorly equipped to process large numbers—even for those engaged in scientific research (Drane et al., 2009; Resnick et al., 2017). So, let us do our best to put them in terms that make them even somewhat comprehensible.

When we multiply the number of firing patterns in the brain at any given moment (10 to the millionth power) with the approximate number of chemical reactions occurring throughout an average lifespan (10,000 per second, roughly 25 trillion), the resulting astronomically surpasses the number of particles in the known universe. It exceeds the limits of traditional mathematical notation and current computational power, rendering it impractical to express or calculate.[74]

Consider this in terms of the constraints and nature of some of our present technologies. Notwithstanding temporal and spatial resolution worries, the mean voxel size for an fMRI scan ranges from 1 to 4 mm (Düzel et al., 2015; Uludag et al., 2015). Deep Brain Stimulation (DBS), a relatively precise neural modulation method, utilizes four electrode contacts, each 1mm in length and having a pitch between 0.5-1.10 mm (Butson & McIntyre, 2006; Wu et al., 2021).

For the sake of simplicity, let's presuppose both techniques purport to depict or modulate a neuronal cluster of approximately 2 milligrams—roughly one and a half thousandths of one percent of the entire brain mass. This minuscule segment of organic material, small enough to balance on a pinhead, comprises roughly 15 million neurons. If placed end to end, these neurons would extend for 450 kilometres, creating 150 billion connections in total—a number tenfold the estimated age of the known universe in years. Each could reflect as much as 10,000 chemical reactions per second.[75]

Technologies will continue to improve, as will network analysis and the examination of large-scale connectivity patterns.[76] But even assuming this is the case, the far more pressing issue is the extent to which particular spatial or functional aspects of the brain are 'localizable.' This means that specific mental functions or actions can be linked to the activity or activation of certain brain regions. The problem is many imaging technologies, forms of neuromodulation, and corresponding studies on which they are based are premised, at least in part, on the principle of localization.

A concept known as the "fallacy of localization" pertains to the reductionist proposition that unique mental functions can be attributed solely to distinct, well-defined areas within the brain. Contrary to this oversimplified viewpoint, the human brain should be understood as an intricately interconnected network where multifaceted cognitive functions often necessitate the harmonious interplay among discrete cerebral regions.

This misleading notion predominantly stems from the misinterpretation of neuroimaging studies. Frequently, these studies reveal activity within certain regions as correlated with specific tasks or mental functions, which can lead to the presumption of a one-to-one correspondence between a region and a function. Nonetheless, it is essential to appreciate that our understanding of the brain's function and structure remains complex and should not be reduced to overly simplistic associations.

Certain brain areas, such as those previously discussed, are considered to *some* degree 'localizable,' depending on how we understand the term. Basic sensory and motor functions, like those associated with the visual and auditory cortex, offer examples of partially localizable functions. Yet, even these 'localized' areas cannot be fully understood in isolation from their interconnections, such as environmental features providing information and executive functions processing this information. That said, it is fair to acknowledge that our interventions in

neuroscience are not always predicated on complete understanding or localization. Practices such as general anesthesia and electroconvulsive therapy, which proceed based on observed outcomes rather than complete causal understanding, exemplify a pragmatic approach in medical science.

In acknowledging the practical realities of neuroscience, it is essential to recognize that not all interventions in the field are predicated on a complete understanding of these localizations or their broader network interactions. Practices such as general anesthesia and electroconvulsive therapy are prime examples where interventions proceed based on observed outcomes rather than a full causal understanding. These approaches demonstrate the field's capacity to balance empirical evidence with theoretical models, advocating for a pragmatic approach that prioritizes patient outcomes over complete mechanistic clarity.

However, problems with localization seem to hold, in the most profound sense, with respect to brain regions and vast networks that are thought to be associated with 'morality.' In stark contrast to basic motor functions, the so-called 'moral' or 'criminal' brain is immensely complex. They are also thought to overlap with the DMN and networks associated with consciousness (NCCs), themselves various non-localizable brain regions—the Medial Prefrontal Cortex, PCC, Angular Gyrus, Hippocampal Formation, and Parietal Temporal and Occipital Lobes.

This complexity underscores the need for caution in applying simplistic localization theories to complex cognitive functions. While we utilize localization to guide certain interventions, the ethical landscape of neurointervention demands a nuanced understanding of brain function that respects the intricate interplay of various brain regions. As we navigate these complexities, it becomes crucial to continually reassess our approaches and ensure that they are supported by both empirical evidence and a robust ethical framework.

While contemplating the vast convolutions and functionalities of the universe, much greater than those of the particles that weave its very fabric, consider also the lifetime evolution of synaptic connectivity, governed by the same laws as those same particulars which at a quantum level frequently display non-deterministic complex behaviour (Atmanspacher, 2020; Tegmark, 2000; Tse, 2018). Factor in the rise of self-organizing properties, the reciprocity of causation, and the nuances of social and relational dynamics. Understand the brain in the wider context of societal and normative dialogue that both influences and is, in turn, influenced by the brain itself. Weave into this intricate puzzle the concept of mental time travel, the temporal properties of

mental states, and the mysteries surrounding phenomenological experience and human consciousness.

At this exact instant, as you ponder over these issues, know that the very object of your contemplation—the brain—is a biological entity that, were it possible, could decode these enigmatic subjects by deploying the very mechanisms under consideration.

Further, the complexity of a phenomenon should not be equated with its incomprehensibility. While the human brain is undoubtedly a marvel of complexity, ongoing research and advancements in neuroscience reveal that its intricacies can be systematically explored and comprehended, to a limited degree—through various methods, such as network analysis and the examination of large-scale connectivity patterns (Bassett & Gazzaniga, 2011; Heylighen et al., 2006; Sporns et al., 2005). This includes complex phenomena which require taking into account the emergent properties that arise from the interactions and relationships between simpler components.

Moreover, history has taught us human ingenuity and our remarkable powers for innovation and understanding. By employing interdisciplinary approaches and leveraging our cognitive faculties, we continue to unravel the mysteries of the brain, shedding light on its structure, functions, and the remarkable nature of human cognition.

But that said, 'everything should be made as simple as possible, but not simpler.'[77] While we strive for simplicity in explanations, we should also acknowledge the inherent complexity of phenomena and the limits of our own understanding. The question I end with, and that I pose to those who seek to rely on assumptions about 'safety and efficacy' is this: where, then, in this three-pound mass that we call the brain, the universe within and that into which it extends, and the emergent mind, through vast networks spanning space and time, do we locate fundamental aspects of our nature—morality, criminality, rationality, and all they entail. And in turn, to truly find they are 'safe and effective,' how vast are the assumptions required? And how safe is 'safe enough'—particularly given implementational concerns in punishment institutions with a history of human rights violations against the vulnerable and marginalized. I call for any discussions in the realm of ideal theory that stand to theorize about technologies deployed to alter 'morality,' turn a careful mind to these issues, and, with epistemic humility, lay bare the magnitude of these assumptions when addressing ethical issues in the realm of ideal theory.

Neurointerventions—Conceptual Challenges

Even if we set aside the plethora of concerns about practical implementation and assume an omniscient knowledge of the human brain and mind and technologies that are 'safe and efficacious,' I still think moving up the spectrum and approaching the realm of ideal theorizing, there remain challenges, which I only briefly highlight here.

The Scope and Limits of Moral Cognition: Exploring the Relationship between Ought and Is in the Brain

A critical consideration in discussing the nonconsensual use of neurointerventions for criminal offenders is the careful examination of what it means to label a particular neurointervention as 'safe' or 'effective.' However, sustaining a clear distinction between these concepts might pose challenges, particularly when navigating the boundaries of ideal and non-ideal theory. This is because notions it is at least arguable, 'safety' and 'efficacy' tend to incorporate normative dimensions, which demand careful scrutiny.

The concept of effectiveness in the context of neurointerventions appears to be closely intertwined with morality, manifesting in two distinct aspects. First, such interventions aim to target specific regions of the brain associated with moral reasoning and decision-making. However, our current understanding of the neural mechanisms underpinning moral decision-making remains far from comprehensive, and it is unlikely that we will achieve a comprehensive understanding in the near or distant future.

Even if we were to achieve substantial progress in this area, the concept of 'effectiveness' might be seen to encompass additional moral dimensions. For instance, in an initial study involving fMRI and ethical dilemmas, Greene and colleagues concluded that their findings addressed a "psychological puzzle, not a philosophical one." They emphasized that their study did not claim to determine actions as morally right or wrong (J. D. Greene et al., 2001, p. 2107). This observation holds significant weight since attempting to do so could potentially fall into the trap of committing the 'naturalistic fallacy' (Moore, 1903).[78] It is crucial to recognize that vibrant brain images cannot uncover the regions responsible for embodying moral principles such as the 'golden rule' or the pursuit of overall goodness, nor can they vindicate the truth of moral claims.

But to deem a specific neurointervention effective in addressing immoral behaviour, it seems, at first glance, that we require some sort of 'yardstick' for measuring what is morally right

or wrong. However, just as there is no consensus on the correct scientific theory of moral reasoning, there is also no consensus on a comprehensive theory of morality—a systematic approach to understanding what is right or wrong, good or bad, and how individuals ought to behave in moral situations.

The study of morality remains one of the most divisive and rich fields of inquiry in both ancient and contemporary philosophy, giving rise to fervent debates among diverse disciplines and social groups. This underscores the so-called 'yardstick objection' that has been debated in the context of punishment equivalence and related discussions of moral enhancement (DeGrazia, 2014, p. 364; Earp et al., 2018, p. 168; Ryberg, 2020, p. 57; R. Sparrow, 2014, p. 22). As Sparrow frames it, "If we are going to start giving people drugs to make them more moral, we had better know what it is for someone to be more moral' (DeGrazia, 2014, p. 364; Earp et al., 2018, p. 168; Ryberg, 2020, p. 57; R. Sparrow, 2014, p. 22).

This concern is amplified by the fact that because the mind is extended and embedded, assessing a particular action of disposition as moral or immoral requires considering the conditions of the immediate physical or social environment. Just as the mind cannot be reduced to the brain, neither can the circumstances that explain moral judgements about what is right and what is wrong, what we ought to do, and what we owe to others. This depends on complex facts about circumstances and events in the physical and social world in which such matters are situated.

One response to this dilemma is to argue that for an intervention to be considered 'effective,' it would be sufficient to increase an offender's 'capacity' for moral reasoning. Such an enhancement could facilitate the individual to act morally rather than simply controlling their behaviour by dictating a particular course of action (Earp et al., 2017; Earp et al., 2018, p. 167; Focquaert & Schermer, 2015; Lewis, 2021, p. 18; McMillan, 2014; Raus et al., 2014).

However, it is important to note that an increased capacity for moral reasoning does not guarantee that an offender will consistently act morally; in fact, the opposite may be the case. It is difficult to see how this could be seen as 'effective' in all cases. These potential concerns regarding the consequences of increased capacity for moral reasoning will be explored in later chapters, specifically focusing on the potential threats that neurointerventions pose to human freedom.

Another response would be to deem neurointervention 'effective' based on its ability to prevent or reduce crime, specifically in relation to actions that are prohibited by law. This seems intuitive because the focus is punishment equivalence. However, it is crucial to recognize that there is a distinction between what is considered moral and what is legal. The definition of criminal acts is a result of political decision-making, which can change over time (Ryberg, 2020, pp. 14-15). This issue ties into the problem of overcriminalization; as highlighted in the previous chapter, it is identified as a significant factor contributing to overincarceration.

An illustrative example of the impact of overcriminalization is the case of the criminalization of consensual same-sex acts in the 20th century. Alan Turing, renowned for his role in breaking the Enigma Code during World War II and his contributions to theoretical computer science, endured the administration of antiandrogen hormones as a form of 'punishment' for his alleged contravention. This treatment inflicted severe harm on his physical and mental well-being before his tragic death by suicide in 1954 (Hodges, 1983; Liberto, 2018, p. 196; McTernan, 2018a; Ryan, 2020, p. 272).[79]

So it seems if punishment equivalency arguments want to rely on the assumption that a particular neurointervention is *effective*, they are committed to making at least some normative judgements, at least in a certain class of cases. If the desired effect of administering antiandrogen hormones in Alan Turing's case was preventing him from engaging in consensual same-sex acts, and it achieved this result, we could certainly say it was 'effective' in 'treating crime.' But, at least intuitively, this does not seem satisfactory. In the context of any project for the use of neurointerventions as tools for punishment or crime prevention, I think more needs to be said, even in the realm of ideal theory.

It is valid to recognize that the lack of a definitive 'yardstick' for moral behaviour is not exclusively problematic for neurointerventions but is a challenge across the criminal justice system. Variability in criminalization reflects shifting moral judgments, underscoring a general critique of the assumptions underlying all forms of punishment, not just those involving neurointerventions.[80]

However, while this critique applies broadly, neurointerventions may introduce specific complexities due to their direct and potentially irreversible alterations to brain function. For example, the hypothetical treatment applied to Alan Turing raises particular concerns about permanence and autonomy—issues that might not be as pronounced with more conventional

punitive methods. This highlights the need for further exploration into how such interventions uniquely interact with foundational ethical principles. Acknowledging these nuances is crucial to critically assess the assumptions of safety and efficacy that support the use of neurointerventions in criminal justice—and, in turn, identify further areas of theorizing for punishment equivalence.

As a final note, similar conceptual concerns arise regarding the assumptions surrounding the 'safety' of interventions. Concerns about 'safety' could be understood in different ways. For example, one might take 'safety' to involve an inquiry into broader considerations related to a specific population in a given context, including risks to the safety of vulnerable prison populations and the potential perpetuation of inhumane practices despite lessons from history. Moreover, such assumptions may jeopardize collective values upheld in Western liberal traditions, which are recognized as a dynamic extension of a community's different conceptions of morality. In this sense, to assume an intervention was 'safe,' in a general sense, requires assumptions which may not always be the case.

While not explicitly stated in the prevailing discussions of punishment, I think it is charitable to assume theorizing under the 'in-principle' constraint tends to narrow the interpretation of 'safety' to potential medical side-effects on the 'deviant other,' with a specific focus on immediate risks, particularly pertaining to the brain. I leave open the issue of whether equivalency arguments might rely on the 'in principle' constraint to justify setting these matters aside—subject to the qualification that such assumptions should be laid bare if such arguments are intended to inform practical implementation. I set this aside now, and in the final chapter of this dissertation, I revisit and explore practical and contingent considerations. These considerations address the recognition of rights and the protection of the mind, extending these principles to all individuals, including prisoners.

Conclusion and Final Thoughts

In closing, as we strive to unravel the mysterious workings of the human mind, it is essential to recognize the constraints of our present understanding. While the exercise in this chapter has been tedious, it will become clear why many of these matters currently, and through future research, open valuable avenues of inquiry related to punishment, equivalence, parity, normative concerns, and rights pertaining to mental processes—illuminated unresolved puzzles that are not always acknowledged.

The convergence of various technologies, such as neuroimaging, neuromodulation, and computational neuroscience, showcases the remarkable synthesis driven by human ingenuity. These advancements hold promise for improving the quality of life in the clinical setting and warrant consideration in other contexts—including criminal justice practices.

I conclude by stressing caution must be exercised when venturing beyond clinical applications and delving into the territory of altering distinct and highly complex human features. In light of this, ethical theorizing plays a crucial role in addressing immediate and foreseeable issues posed by potential interventions to alter human morality and treat socially undesirable behaviours. This is where ethical theorizing is of crucial importance. And the need for such theorizing is widely acknowledged to address specific issues about the appropriate use of neuroscience in the criminal justice system.

Alive to the limits of our current understanding and possible comprehension, what is required is a 'slow science,' which encourages critical reflection on the means and ends of scientific endeavours; ethical theorizing in the context of neurointerventions should embody a methodological approach that values careful consideration and improvement of the world we live in (Baylis, 2019, pp. 124-125). While avoiding concerns about 'status quo bias,' in what follows, I argue the current limitations of our technologies support the recognition of a refined version of a 'precautionary principle' to address both ethical and legal issues well in advance of the refinement of existing technologies, and development of novel technologies (Caney, 2009; McKinnon, 2012, p. 56; Resnik, 2003; Resnik, 2004; Sandin, 2009).

The role of ideal theorizing continues to hold significant value. Yet, similar to the 'baseline objection' and the practical issues that emerge in reality, worry about safety and effectiveness, which are magnified by our severely constrained understanding of the human brain, present substantial hurdles to arguments advocating punishment equivalence within the sphere of non-ideal theorizing. This focus on the state of affairs in the real world suggests that such concerns are likely to persist into the foreseeable future.

Even as we venture to stretch the parameters of our imagination, we must concurrently acknowledge the limitations of our current technologies. Simultaneously, we must maintain a mindful cognizance of the boundaries of our own rationality and the potential risk that they may obscure our capacity to draw on the higher echelons of our nature as we strive to confront crime and diligently seek evidence-based solutions in a manner that is both responsible and humane.

We must stand prepared, for without readiness, our ability to adapt and theorize may falter, unable to keep pace with the rapid progression of technology. For while, at least at present, we remain constrained by the limitations of our understanding and corresponding risks, the universe—from which we, its subjects, arise—is not obliged to respect those bounds.

3

The Principle of Parity

Widespread Subversion, the Inner Sphere, and the Bounds of Human Rationality

Introduction

In the recent decade, Wim Hof, a vibrant Dutch national of sixty years, has captivated the media spotlight due to his astounding ability to endure lengthy and regular exposures to severe cold. His apparently superhuman feats of resilience have taken many forms: prolonged submersion in frigid Arctic waters, running a half marathon barefoot above the Arctic Circle and achieving a world record for standing encased in a container filled with ice for nearly two hours. These remarkable accomplishments have seen him bestowed with the epithet 'Iceman.'

Hof credits these feats to developing a technique involving a combination of forced breathing, cold exposure and meditation—which he has trained and taught—known as the 'Wim Hof Method' (Hof, 2020; Morris, 2021). Hof claims that developing and practising this method lets a person exercise conscious control over the operation of the cardiovascular system, including heart rate, blood pressure and body temperature.

Understandably, Hof's feats have been met with a heavy dose of skepticism. This is because the brain mechanisms governing these processes primarily reside in the deeper primitive recesses of the brain, far down the cortico-subcortical hierarchy. They function as primitive life-sustaining mechanisms and are generally seen to operate autonomously. How, as Hof does, could one reasonably claim to exert 'conscious control' over these processes?

In turn, these feats have attracted the attention of the scientific community. Over the past years, researchers have conducted multi-modal imaging studies testing these claims. (Muzik et al., 2018). The results were surprising, and we will explore them to motivate challenges to conventional views on conscious control and human rationality later in this Chapter, as we discuss the parity principle.

The 'parity principle' proposed by Neil Levy (2007, 2020) has emerged as a pillar of punishment equivalence arguments. It contends that, in the absence of ethically relevant differences, we should treat direct neurointerventions on par with traditional methods of

influencing the mind. If such neurointerventions are safe and effective, no ethically significant distinction exists between these novel methods and more conventional ways of changing minds. Consequently, our focus should be on the *effects* of these interventions rather than on the specific *means* by which they are achieved. In this chapter, I identify what I see as certain challenges to the parity principle.

One of the more compelling responses is that the means do matter. This is because neurointerventions have the unique ability to bypass 'rational capacities' and 'conscious control' that traditional ways of changing minds do not. Because they are 'freedom incompatible,' there is a 'normative asymmetry' (Bublitz, 2020a, 2020b; Bublitz & Merkel, 2014; Focquaert & Schermer, 2015; Harris, 2011b, 2012, 2013, 2014b, p. 372; Shaw, 2014).

In response, those who defend the parity principle have argued this is not a property unique to neurointerventions, pointing to a phenomenon I describe as 'unconscious subversion' (Sunstein, 2014; Sunstein & Thaler, 2003; Till, 2012). The fact is that many forces in the environment and conventional means we use to change the minds of others also circumvent our 'rational capacities' and 'conscious control' (Levy, 2020). This is due in part to the limits of human rationality, which we explored in previous chapters (Davies, 2009, 2020; Kahneman, 2011; Sunstein, 2014; Sunstein & Thaler, 2003; Till, 2012; Wilson, 2002).

In this chapter, I explore the debate surrounding direct and indirect interventions within the context of criminal justice issues. However, I believe that the concepts of 'rational capacities' and 'conscious control' do not accurately reflect the complexities of our mental lives and human rationality. The current dichotomy neglects the importance of our ability to self-regulate our thought processes. Moreover, the parity principle often overlooks significant aspects of human rationality, such as the influence of our physical, social, and normative environments on rational agency.

Although these dilemmas pose challenges for the parity principle, they are not insurmountable barriers. However, they do highlight the need for additional theorizing, particularly in regard to the unique challenges posed by neurointerventions compared to traditional modalities like imprisonment. We must be cautious about altering anything that might jeopardize a significant form of human freedom or intrude upon a sensitive, valuable 'inner sphere.' This dialogue sets the stage for the concluding chapters, which will investigate potential

threats of neurointerventions in greater detail, including the possibility that direct interventions pose unique threats to 'mental freedom'.

The Parity Principle—Direct and Indirect Interventions

In the last chapter, we delved into the rise of neuroscience innovations. In contemporary discourse, in a broader debate about 'moral neuroenhancement,' discussion surpasses the confines of criminality, addressing the broader implications, both the promise and pitfalls, of enhancing human abilities in various fields and broader societal efforts to prevent grave harm or extinction (Buchanan, 2011; Carman, 2021; Earp et al., 2018; Focquaert & Schermer, 2015, p. 139; Harris, 2011, 2014a; Persson & Savulescu, 2008, 2011a, 2011b, 2012, 2013, 2015; Savulescu & Persson, 2012; Sparrow, 2013; R. Sparrow, 2014; R. J. Sparrow, 2014; Wiseman, 2016).

The application of new technologies to alter our core nature, especially at the societal level, triggers deep unease. This worry stems from the perceived hubris of 'playing god,' disrupting our unique universal connection and potentially leading to synthetically created moral agents. It also stokes fears of a looming 'posthuman future' (Erler, 2020; Fukuyama, 2003; Habermas, 2003).

The forced use of these technologies in criminal justice intensifies this discomfort, as it seems to assault human dignity, reducing offenders to 'machines' needing repair and possibly destabilizing societal structure (Bomann-Larsen, 2013; Lavazza, 2017; Shaw, 2018; R. J. Sparrow, 2014).

The primary concern is the threat these interventions pose to 'human freedom,' an ideal that lets us live 'unfettered by the given.' The fear is that such technologies could undermine a 'self-determined life'—the freedom to shape our own lives—which forms the essence of moral virtue and perhaps morality itself (Harris, 2011, 2014b; J. Harris, 2016; Kant, 1999 (1781); Locke, 1824a).

We will survey and sort through some of these concerns in the next chapter. Nonetheless, thorough and ethical analysis requires us to go beyond mere intuition and alluring rhetoric. Certain views in the public discourse exhibit aversion towards the 'novel, neuro, and the nonnatural', contributing to the emergence of a 'status quo bias'. This bias favours the

preservation of the current state of affairs, frequently at the expense of reasoned debate and innovation (Bostrom & Ord, 2006; Bublitz, 2020b; Caviola et al., 2014; Dresler et al., 2019, p. 1142; Persson & Savulescu, 2012, p. 115).

Enter Levy's parity principle. As Levy explains: "there seems to be a widespread presumption in favour of traditional ways of changing minds, other things being equal" (Levy, 2007, p. 71). As Levy asks: "Why is the newer technology so much more controversial than the old? Is there a fundamental difference between different kinds of interventions, and is that difference genuinely morally significant?" (Levy, 2020, pp. 34-35). The parity principle seeks to address this issue.

Direct and Indirect Interventions

Levy's parity principle rests on an essential distinction, asserting that there exist "two basic ways to go about changing someone's mind" (Levy, 2007, p. 69)—a proposition I will dispute shortly.

The first category encompasses *direct interventions*,[81] also known as 'neurointerventions'. These interventions include chemical, electrical, and surgical procedures that act directly on the brain—such as psychopharmacological methods and neuromodulation. For instance, the administration of an SSRI or the modulation of brain activity through tDCS are examples of attempting to directly alter the brain—for example, to curb undesirable behaviours, such as violent aggression.

The second category covers *indirect interventions*, which involve traditional methods of influencing others' minds. These methods often involve changes to the physical or social environment and interpersonal interactions. For instance, in a criminal context, confining an offender in prison would involve altering their physical and social environment. Requiring offenders to engage in cognitive behavioural therapy (CBT), such as 'anger management' classes, is another example of attempting to change their minds indirectly.[82]

The Extended Mind

The parity principle builds on a particular philosophical theory known as the Extended Mind Thesis (EMT). The EMT proposes that the brain, though the primary 'vessel' of the mind, is not its exclusive realm. Rather, the mind extends outward, embracing objects within our physical environment (Clark & Chalmers, 1998; Damasio, 1994; Heersmink, 2016; Heinrichs,

2018, p. 60; Hurley, 1998). In this context, Levy describes the mind as 'supervening' on the brain, which is to say, "there are no mental differences without corresponding neural differences" (Levy, 2007, p. 62; 2020, pp. 34-35). Put another way, the mind can be 'co-realised by carriers beyond the brain' (Heersmink, 2016; Heinrichs, 2018, p. 60).[83] This aligns with our previous discussion, which highlighted the mind as an entity that extends and embeds itself into a wider physical and social environment, participating in continuous cycles of action and perception.

Direct, Indirect, Means and Effects

If we accept the EMT, both direct and indirect interventions could, in essence, lead to similar outcomes—they influence the brain's operations, or "what neurons fire when and how" (Greely, 2008, p. 1134; Levy, 2007, p. 62; 2020, pp. 34-35). For example, the administration of an SSRI or the direct modulation of brain activity via tDCS can change the brain's information processing and energy patterns, altering the mind. The prison environment, particularly extended periods of isolation, can significantly impact the mind and thus cause changes in the brain. In both instances, the brain and mind are subject to change. This raises the question: What, if any, is the substantial difference between these approaches? And does this difference bear ethical relevance?

Ethically Relevant Differences

Levy argues the parity principle entails that unless we can identify *ethically relevant differences* between direct and indirect interventions and alterations, we ought to treat them on par (Levy, 2007, p. 62). In technical terms, Levy states:

> Alterations of external props are (ceteris paribus) ethically on par with alterations of the brain, to the precise extent to which our reasons for finding alterations of the brain problematic are transferable to alterations of the environment in which it is embedded (Heinrichs, 2018, p. 63; Levy, 2007).[84]

Put another way, this principle is based on the idea that our concerns about altering the brain should apply equally to altering the environment in which the brain operates. *Provided other conditions remain the same*, if we have ethical reservations about one, according to this principle, we should have similar reservations about the other, as both can have profound effects on the mind and behaviour.[85]

The crux of the matter is whether direct brain interventions carry an inherent ethical objection that sets them apart from conventional interventions, such as incarceration. Given that both direct and indirect interventions produce the same effect, and the means do not matter, this serves to challenge some of the more problematic intuitions we have about intrinsic properties of neurointerventions 'novel, neuro, and nonnatural.'

Circumventing Rationality, Subverting Freedom

The parity principle has recently faced scrutiny from multiple theorists, suggesting that direct interventions in the mind and brain, especially those employing advanced neurotechnology, deviate from indirect methods of mental alteration. This proposition, known as the 'Asymmetry Claim,' posits an ethically important difference between direct and indirect interventions (Bublitz, 2020a, 2020b; Bublitz & Merkel, 2014; Focquaert & Schermer, 2015; Harris, 2011, 2012, 2013b; 2014b, p. 372; Shaw, 2012).

The most compelling argument in support of the claim suggests that direct interventions possess a distinct characteristic absent in indirect interventions. Specifically, direct interventions have the ability to 'circumvent rational capacities' and bring about changes 'without the conscious awareness' or 'active engagement' of the subject (Bublitz, 2020a, 2020b; Bublitz & Merkel, 2014; Focquaert & Schermer, 2015; Harris, 2011, 2012, 2013b; 2014b, p. 372; Shaw, 2012). In this sense, it is argued the means do matter.

The parity principle categorizes interventions as 'direct' or 'indirect'. Yet, in discussions about 'circumventing rationality,' theorists differentiate further between 'active' and 'passive' interventions. '*Active interventions*' demand the recipient's active involvement and effort, while '*passive interventions*' apply their effects without such active input (Bublitz, 2018, p. 291; Bublitz & Merkel, 2014, p. 69; Focquaert & Schermer, 2015, p. 28; Raus et al., 2014; see also Schermer, 2015). This active/passive dichotomy hinges on a general concept of 'rational capacities', prompting some scholars to label them, correspondingly, as 'rational' and 'arational' interventions (Douglas, 2018, p. 215)—and I will do so here.

As we delve deeper, it's paramount to thoroughly examine terms like 'circumventing rational capacities' and 'conscious control' and the implications neurointerventions may have for them. Theorists have framed these terms in different ways when considering neurointerventions. As above, some discuss the absence of 'active involvement and effort' (Bublitz, 2018, p. 291;

Bublitz & Merkel, 2014, p. 69; Focquaert & Schermer, 2015, p. 28; Raus et al., 2014; see also Schermer, 2015). Others discuss this in terms of bypassing 'dilatory processes' such as 'thought or reflection' and acting 'directly on the mainsprings of action, on emotions or other dispositions' (Harris, 2012, p. 294). Some discuss the potential for direct interventions to bypass "psychological (not necessarily rational) processes altogether" and even "change the cognitive machinery itself" (Bublitz & Merkel, 2014, p. 70; Craig, 2016, p. 115). More refined accounts focus on how direct interventions reach the brain through 'nonperceptual routes', with reference to the complexities of 'conscious control' within the context of the dual network theory—system one and system two processing (Bublitz, 2020b, p. 58).

Most, if not all, of these accounts seem expressly or implicitly to assume, perhaps not unreasonably, that rational engagement requires conscious control—at least to some *degree* (Bublitz, 2020b, p. 65). It has recently been noted that the relationship between conscious awareness and control does not always align; individuals may be aware of an intervention's effects yet unable to resist them.[86] This adds an important insight to the debate. Nonetheless, notions like 'rational capacities' and 'conscious control' are intricate, and their specific interpretations are based on a variety of foundational theories, principles, and related empirical assumptions, which create a web of conceptual complexities.

If direct interventions are a-rational—they circumvent rational capacities—then this seems, at first sight, to be a difference in the *means* by which they change the brain—not simply the effect. If this is true, then the question would be if this is a *morally relevant* difference. One argument is that by circumventing rational capacities, neurointerventions are "freedom subverting" (Harris, 2014b, p. 372), and this matters morally—for any number of reasons. This is a topic for the next chapter when we consider the 'freedom objection.'

But for now, if we reasonably accept neurointerventions bypass rational capacities—as I think we can—the question is whether this is an *in principle* difference. Whether this is a distinctive property that neurointerventions have. But when we look closer, we see this is not necessarily the case.

Widespread Indirect Subversion and the Limits of Human Rationality

To summarize, the parity principle outlines two fundamental methods for modifying the brain: 'direct interventions' and 'indirect interventions.' Scholars have introduced another layer

of differentiation, separating interventions into 'rational' and 'arational' categories. 'Rational' interventions actively engage the individual's 'rational capacities'—involving, on some accounts, active participation and control. On the other hand, 'arational' interventions operate without engaging these capacities.

It appears plausible to consider that neurointerventions are generally 'arational', acting directly on the brain without engaging 'rational capacities'.[87] Suppose we harbour ethical concerns about this aspect. To address the parity principle, we must also question if similar reservations apply to indirect interventions. But despite neurointerventions being 'arational', many 'indirect interventions' also share this characteristic.

When we understand 'indirect interventions' as changes to the physical or social environment, matters are not so straightforward. This grounds many of the more compelling arguments in contemporary literature in defence of the parity principle and centres on a phenomenon known as 'indirect subversion.'

As we have discussed up to this point, the truth is we are not rational in the ways we generally suppose. We are encumbered on every front by a powerful host of cognitive and heuristic biases, leaving us vulnerable to powerful forces that shape our actions and behaviours in profound ways which we cannot fully comprehend. As Levy puts it, indirect subversion is an "all too pervasive and all too powerful feature of the world we live in," and all of us are manipulated on a daily basis in ways to which we do not consent, using indirect interventions" (Levy, 2020, pp. 44-45). For Levy, then, the direct/indirect distinction does 'not map onto anything of ethical significance' nor even serve as a 'useful heuristic' to identifying problematic interventions (Levy, 2020, p. 46).

More significantly, Levy believes that the parity principle underpins the concept of 'externalist ethics.' This notion underscores the intricate ties between individuals and their context, casting doubt on the traditionally rigid delineation between the two. By adopting this viewpoint, we can discern that our apprehensions about direct interventions are inextricably linked to our ability to identify some of the "most dramatic and important injustices in the world" (Levy, 2007; 2020, p. 46; Lippert-Rasmussen, 2018, p. 153).[88]

Levy's use of the parity principle strays somewhat from purely abstract theorizing, taking into account the realities of our world and mental phenomena while still focusing on conceptual issues. This is a useful form of theorizing, which has led to comprehensive rebuttals and what I

see to be more dynamic and sophisticated philosophical exchanges (Bublitz, 2020b; Levy, 2007; 2020, p. 46; Lippert-Rasmussen, 2018, p. 153). These further discussions about 'indirect subversion,' boundaries between agents, and the scope and limits of our rationality serve as a vital entry point for framing issues to follow.

Strangers to Ourselves—the Limits of Human Rationality[89]

In everyday life, and sometimes even in philosophical discussions, we perceive ourselves as free, rational agents with conscious control, driven by beliefs, desires, and intentions. Yet, as we will see, the reality is quite different. This is not the types of creatures we are—although we may aspire to become so.

The Adaptive Unconscious

I want to return to the discussion of 'consciousness' we discussed in the last chapter. The concept of the unconscious as an "unknown," "imperfectly reported" substratum sending forth "vapours, odd beings, terrors, and deluding images" imposing a powerful influence on our lives is neither novel nor astonishing (Campbell, 2008; Freud, 1899 [1965], p. 198; Jung, 1959).[90] However, alongside significant advances in neuroscience and an increasingly comprehensive biological understanding of its origins, there is a growing appreciation of just how profound its impact is on our lives.

The prior chapter laid the groundwork for the Global Workspace Theory (GWT), a paradigm that describes 'consciousness-as-reportability,' fueled by self-sustaining signals that proliferate extensively across a network within the cortical region (Baars, 1997; Dehaene & Naccache, 2001; Mashour et al., 2020). Despite this, a substantial segment of our mental processes, the actual foundations of our actions, fail to reach ignition within this workspace, and remains 'silent, hidden from our conscious selves" (Davies, 2009, 2020; Kahneman, 2011, p. 52; Ryberg, 2020, p. 90; Wilson, 2002).

There are many reasons why this may be the case. One theory is that this is related to the brain's architectural structure, specifically the cortico-subcortical hierarchy we have considered. Nestled deep in the ancestral mammalian brain, alongside primitive life-sustaining mechanisms, are thought to be "built-in" affective submodules dedicated to discriminatory, emotional, and social capacities. These endogenous—internally generated—submodules function

autonomously, evoking powerful emotion (Bechtel, 2007; Davies, 2020, p. 325; Panksepp, 2004, 2012).[91]

For example, children born without cortical neurons, responsible for higher-level cognitive functions, still display emotions and engage in social behaviour (Davies, 2020, p. 329; Merker, 2007; Panksepp, 2004; Shewmon et al., 1999, p. 371). Similar emotional submodules are not exclusive to humans but are shared across various mammalian species. For instance, a young rat instinctively freezes in fear when encountering cat fur for the first time despite lacking prior exposure to cats (Panksepp, 2012, p. Chapter 5). This is significant because at least *some* ancestral submodules, like physiological states, operate independently of conscious awareness and corrections are simply 'absent from our neural architecture' (Davies, 2020, p. 344).[92] Strong emotions like anger, fear, or disgust significantly influence decisions and actions, and this information does not directly reach conscious awareness.

Even when neural structures permit it, signals may not reach our awareness due to an information overload or weak signals failing to assert themselves. Further, our attention, which can be overwhelmed by incoming signals, might neglect some information, especially if it comes from weak or sporadic sources. Moreover, even when signals reach consciousness, interpreting the resultant physical and emotional responses is not always accurate; these interpretations are "demonstrably prone to error" (Davies, 2020, p. 331).

Signals from our ancestral brain significantly influence our decisions, actions, and failures to act in ways of which we are not consciously aware. This is due to our interpretative limitations, weak signal strength, and missing connections in our neural architecture. So, we humans, viewing ourselves as rational beings, tend to rationalize our actions, oblivious to their true origins—a phenomenon known as 'confabulation' (Davies, 2020; Gazzaniga, 2000; Maier, 2020).

'Mental Shortcuts'—Cognitive and Heuristic Biases'

These same sub-cortical signals, which often elude our conscious control, are those that were evolutionarily adapted by humans and mammalian ancestors to prioritize satisfying and easy-to-achieve immediate needs necessary for survival—such as finding food, shelter, and avoiding danger. This is confirmed by various neuroimaging studies, which indicate our limited

capacity to conceptualize ourselves beyond immediate needs or the near future (Davies, 2020, p. 339; Mitchell et al., 2011; Pronin et al., 2008; Pronin & Ross, 2006; Wagner et al., 2012).

However, while essential for survival, certain brain mechanisms can restrict our capacity for intricate or long-term thought, which requires significant brain resources. They reflect 'System 1' reasoning, characterized by quick, heuristic-based responses.

Heuristics serve as cognitive expedients, facilitating quick decision-making and problem-solving without constant deliberation. Although efficient and useful in many cases, such reasoning may not yield optimal decisions in complex or unfamiliar scenarios (Banja, 2018, p. 287; Bublitz, 2020b, p. 64; Greene, 2013, p. 133; Haidt, 2001; Kahneman, 2011). This can induce cognitive biases.

Confirmation bias epitomizes this, as individuals tend to seek, interpret, and remember information confirming their pre-existing beliefs (Wason, 1960). The availability heuristic relies on immediate, easily recallable examples when evaluating a concept or decision (Tversky & Kahneman, 1973). The anchoring bias results in an overreliance on a single piece of information during decision-making (Kahneman et al., 1982). Hindsight bias, or the 'knew-it-all-along' effect, denotes our propensity to view past events as predictable (Roese & Vohs, 2012). Stereotyping, an instance of 'implicit bias,' shows individuals making decisions based on stereotypic attributions, even when they explicitly reject these stereotypes, often justifying their choices through confabulation (Holroyd et al., 2017; Moles, 2014; Pearson et al., 2009; Uhlmann & Cohen, 2005)[93].

When we view the mind as continually adaptive, shaped by reciprocal self-organizing processes over time, we recognize how powerful affective forces, limited conscious awareness, and various biases not only influence immediate decisions but also collectively shape our beliefs, dispositions, personalities, and the way we mentally construct reality.

The Veiled Machinations: Unmasking Unconscious Subversion

The limitations of our rationality, which restrict our abilities to process information and understand the origins of our actions, render us vulnerable to 'widespread unconscious subversion,' a phenomenon that can be categorized as 'indirect intervention' for the purpose of maintaining parity in reasoning.[94] It is pervasive, and it comes in various forms, all of which have been discussed in theorizing about the parity principle.

Nudging

In neuroscience, 'nudging' involves leveraging our understanding of the brain's cognitive processes to influence behaviour. It lays prey to cognitive biases and heuristics, exploiting tendencies and subtly altering choice architecture through environmental cues to encourage specific behavioural outcomes. Whether and under what circumstances nudging is ethically permissible is heavily debated (Dolan et al., 2010; Hansen & Jespersen, 2013; Hertwig & Grune-Yanoff, 2017; Reisch & Sunstein, 2016; Sunstein, 2017; Sunstein & Thaler, 2003). In some instances, advocates justify nudges when they are thoughtfully and precisely applied, as they can counteract biases and enhance decisions regarding health, welfare, and happiness while preserving freedom of choice (Thaler & Sunstein, 2008; Viale, 2022). But this is not always the case.

Dark nudges, also known as manipulative nudges, deviate from the traditional aim of nudges by utilizing deceptive or manipulative techniques. Unlike standard nudges that guide individuals towards beneficial choices while preserving freedom of choice, dark nudges exploit cognitive biases for manipulative purposes (Viale, 2022, pp. 1-5, 157-160; Wilkinson, 2012). For example, a leading CIA analyst is reported to have endorsed literature on priming techniques and nudging intelligence officers to utilize tools of deception (Viale, 2022, p. 1). As early as 2012, Facebook conducted an emotional priming experiment involving over 700,000 users, while in 2015, Amazon mentioned a program using machine learning to generate 70 million users (Viale, 2022). The explosion of social media and its role in our lives has raised concerns about the use of nudges on these platforms, including their potential for advertising, digital mass persuasion, and political influence (Bongiovi, 2019; Matz et al., 2017; Tufekci, 2014), compounded by parallel concerns about the social implications of the advancing field of artificial intelligence (AI) and how they could further bolster these tools (Brecker et al., 2023; Illia et al., 2023; Lund et al., 2023).

Unsurprisingly, the surreptitious use of nudges for ulterior purposes has 'renewed currency' of old worries that were initially sparked by propaganda and subliminal advertising, and the precise threats associated with comprehending the neural mechanisms of decision-making remain largely unknown but of pressing concern in contemporary society (Spence, 2020; Stanton et al., 2016).

Priming

Another form of unconscious subversion is known as priming, which takes various forms and involves exposing individuals to certain stimuli or cues that can activate specific concepts or associations in their minds that operate below the level of conscious awareness (Bargh & Chartrand, 1999; Dehaene & Changeux, 2011; Kiefer et al., 2012).

Visual stimuli non-consciously influence judgments and actions, a phenomenon that has been established in a long line of studies frequently discussed in the neuroethical debate surrounding neurointerventions.

For example, in one experiment, subjects subliminally exposed to happy or angry faces rated and consumed a fruit beverage differently. Interestingly, conscious feelings reported by participants were indistinguishable between the happy and angry face conditions (Winkielman et al., 2005). In another experiment, the presentation of a pair of eyes unconsciously influenced subjects' contribution to an honesty box for coffee milk (Bateson et al., 2006; also cited in Kahneman, 2011; Nettle et al., 2013; cited in Ryberg, 2020, p. 90). Unpleasant smells influenced harsher moral judgments (cited in Ryberg, 2020, p. 90; Schnall et al., 2008), and a bitter taste in the mouth led to the perception of different types of ethical behaviour as more wrong (Eskine et al., 2011; cited in Ryberg, 2020, p. 90). Further studies revealed a decrease in favourable rulings by judges before daily food breaks, known as the "hungry judge" effect (Danziger et al., 2011). Priming participants with the stereotype of "elderly" slowed their movements during puzzles (Bargh et al., 1996; discussed in Kahneman, 2011; discussed in Levy, 2020). Participants exposed to polite words exhibited longer interruptions during conversations compared to those exposed to rude or unrelated words (Bargh et al., 1996, p. 230). Finally, painting prison cells with the so-called "Baker-Miller pink" allegedly reduced inmate aggression and incidents of institutional violence, although these findings have been debated (but see also Genschow et al., 2014; Pellegrini et al., 1981; Schauss, 1979).[95] All these instances illustrate how elements within our surroundings influence our behaviour, actions, and decisions. Yet, we remained oblivious to these impacts at a conscious level.

In summary, our brain's architecture obscures many of our actions from conscious awareness, making us susceptible to powerful environmental influences. Nudging and priming exemplify this phenomenon, supported by contemporary neuroscience and behavioural psychology. But I think this has further implications.

When we envisage the mind as an adaptive entity, moulded by reciprocal self-organizing processes over time across temporal, physical and social domains, we appreciate the profound sway of emotional forces, limited conscious cognizance, and an array of biases. These factors not only direct immediate decisions but collectively sculpt our beliefs, dispositions, personalities, and our very construction of reality.

Moreover, recognizing the mind as embedded within a broader social milieu elucidates the manifestation of these influences in collective behaviours—ranging from affinity bias to outgroup dehumanization and even the enduring construct of moral responsibility that informs our criminal justice practices. These social influences, in turn, exert a 'top-down' effect on individual cognition and behaviour and recursively embed these features by perpetuating self-amplifying bidirectional cycles of action, perception, and social construction that often perpetuate injustices, and harm—as is the case, I believe, with our criminal justice practices.

So it is fair to say our standard views of human rationality are often deeply misguided (Davies, 2009, 2020; Wilson, 2002), and widespread unconscious subversion is a pervasive feature of our world (Levy, 2020, pp. 44-45)—far more than we can, perhaps, even comprehend. And while the idea that we are such creatures, with limited access to our minds, and so deeply susceptible to such forces, is surely 'alien to our experience,' 'but it is true' (Kahneman, 2011, p. 52). In the end, we are 'strangers to ourselves' (Wilson, 2002)—both as individual entities and collective organisms, intricately woven into a world and societal structures that remain equally enigmatic. And *perhaps* one might expect it to remain this way.

So, where does this leave us? Levy's parity principle stipulates that our apprehensions around direct brain modifications should be mirrored in concerns about environmentally induced changes. Do the perceived dangers of neurointerventions surpass existing threats in our environment? And how does this affect their potential application in criminal justice? To maximally exhaust all avenues of inquiry, I think it is necessary to dig a little deeper.

The Bounds of Human Rationality—Internal Mentation, Mental Time Travel, and Embedded Mind

To summarize, at its core, the parity principle distinguishes between direct (neurointerventions) and indirect (environmental) interventions. Scholars express concern that direct interventions might 'circumvent rational capacities', thus labelling them as 'arational.' Yet,

it's crucial to understand that both forms of intervention—direct and indirect—can potentially exhibit 'arational' characteristics. In fact, there's an argument suggesting that 'arational' indirect interventions are not just common but pervasively influential, thereby posing substantial risks to human freedom.

This line of reasoning and the debate at large, I believe, presents issues. Here, I confront the parity principle with three challenges. First, I dispute the sustainable distinction between direct and indirect interventions. Second, I underline the overlooked temporal aspects of mental states. Third, I emphasize the often ignored concept of 'embeddedness.' These elements scaffold a broader scope for human rationality by identifying how the very act of discerning its limits could have important implications.

The 'Iceman'—Internal Mentation

Returning to the exceptional case of Wim Hof—the 'Iceman,' famous for enduring extreme cold and allegedly controlling his body temperature and cardiovascular responses consciously—we may now analyze his extraordinary claims.

Typically, our responses to severe cold, such as heart rate and body temperature adjustments, are automatically managed by our autonomic system beyond our *immediate* conscious control—due, in part, to the absence of neural connections with higher cortical regions (Beissner et al., 2013; Critchley & Harrison, 2013; Thayer et al., 2012; Thayer & Lane, 2000).[96]

In 2018, Muzik and colleagues used functional Magnetic Resonance Imaging (fMRI) and Positron Emission Tomography/Computed Tomography (PET/CT) techniques to observe Hof's brain and peripheral nervous system during cold exposure. Interestingly, they detected activity not only in autonomic brain regions but also in cortical areas associated with self-reflection, indicating conscious processing (Muzik et al., 2018, pp. 640-641).

Moreover, Hof's heat generation did not rely on standard mechanisms.[97] The researchers suggest this pattern suggests a potential 'top-down' modulation of autonomic functions, involving "bidirectional interactions between cognitive and affective/homeostatic regulatory networks," which may also provide valuable insights into behavioural modification strategies (Muzik & Diwadkar, 2019, p. 259; Muzik et al., 2018, pp. 640-641).

The discourse surrounding conscious control over autonomic processes bears resonance with our earlier consideration of 'emotional submodules,' which initiate automatic responses that similarly challenge standard assumptions about human rationality (Bechtel, 2007; Davies, 2020, p. 325; Panksepp, 2004, 2012).

However, while Wim Hof's case is fascinating, it is subject to numerous experimental and interpretive limitations, rendering it insufficient to ground strong claims about 'top-down' control through a direct structural link.[98] Navigating these intricate territories of the mind presents us with profound challenges due to the vast complexity of the brain.

I think, at best, it is fair to say Hof's case points towards the need for a more nuanced understanding of 'top-down conscious control,' highlighting a complex interplay among brain processes, homeostatic bodily functions, and environmental interactions, which I expect will be explored alongside ongoing research into Hof's method (Citherlet et al., 2021; Marko et al., 2022; McKinney, 2022; Morris, 2021; Muzik & Diwadkar, 2019; Muzik et al., 2018; Ostermeier & Schmoll, 2022; Petraskova Touskova et al., 2022).

I will leave this work to others. The discussion here, as at all junctures, is intended to stir contemplation regarding the parity principle in light of our current understanding. The issue I wish to raise is, in large part, not empirical but conceptual.

Two Basic Ways to Change Someone's Mind?

Recall the parity principle is premised on the view that there are "two basic ways to go about changing someone's mind" (Levy, 2007, p. 69). In Wim Hof's case, when exposed to extreme cold, his brain, and therefore his mind, was changed. This was reflected in imaging showing activation of brain regions associated with self-reflection and conscious processing.

But if we try to describe what caused this change, I think the parity principle falls short—or at least owed a fuller account. It is not a direct intervention. It does not seem we can describe it as a direct intervention caused *solely* by a change to the physical or social environment impacting the autonomic nervous system. In fact, Hof's exceptional resilience seems to stem from *countering* intense environmental forces—potentially through conscious modulation of associated homeostatic bodily functions. Where does this leave us?

Introspection and Self-Organization

Wim Hof attributes his exceptional capabilities to rigorous training in the 'Wim Hof Method' (WHM), a technique involving forced breathing, cold exposure, and meditation (Hof, 2020; Morris, 2021). It is intriguing to observe similar phenomena in others who adopt similar practices. Drawing from diverse cultural traditions, meditation aims to cultivate a self-regulated attentional state, promoting present-moment awareness, openness, acceptance, and self-reflection (Dahl et al., 2015; Srinivasan, 2019, p. 15; Wallace, 1999; Wallace & Shapiro, 2006; Yakobi et al., 2021).[99] In clinical settings, these techniques are known as 'mindfulness-based interventions' and demonstrate varying degrees of success in addressing certain psychopathologies (Yakobi et al., 2021).

Extensive neuroimaging studies, including comprehensive meta-analyses of meditation practitioners, have consistently shown differences in nerve fibres, specifically in the grey and white matter of the prefrontal cortex, a region strongly associated with cognition, emotion, and top-down control (Chiesa et al., 2011; Fox et al., 2014; Marciniak et al., 2014; Sedlmeier et al., 2012; Tang et al., 2015; Wang et al., 2011).

It seems at least plausible[100] that meditation practices result in consistent and measurable changes to the brain's structure. However, as with Hof's feats, the parity principle appears insufficient in accounting for these effects. Bublitz defines 'indirect interventions' as engaging with perceptual routes, which involve the classic five senses gathering external information (Bublitz, 2020b, p. 58). In contrast, meditation and mindfulness techniques often restrict perceptual input, emphasizing inward attention. So, it appears to seek a satisfactory explanation of what is occurring in Hof's case; we may need to look elsewhere.

Internal Mentation and the Default Mode Network

I think there is an important place to look. In the last chapter, I introduced the concept of self-organization as an emergent property of the mind. Rooted in systems theory and complexity science, reveals the brain's capacity to adapt and reorganize its internal processes through spontaneous activity and connectivity within its neural networks. This captivating aspect of human cognition arises primarily from the brain's internal processes, such as neural connectivity, synaptic plasticity, and network dynamics, operating *independently of external influences*. The

brain's inherent spontaneity and intrinsic dynamics are the primary drivers of this phenomenon (Dresp-Langley, 2020; Kelso, 1995; Szentagothai, 1993).[101]

The default mode network (DMN) plays a crucial role in 'spontaneous internal mentation' and 'stimulus-independent' thought, contributing to self-organization. It comprises a set of interconnected brain regions that are functionally distinct and activated during restful, undistracted states, such as 'mind wandering' or 'daydreaming' (Andrews-Hanna, 2012; Binder et al., 1999; Buckner et al., 2008, p. 20; Svoboda et al., 2006).

The DMN is involved in various cognitive processes, including autobiographical remembering, envisioning the future, generating options, playing out options, and -contemplating decisions. (Buckner et al., 2008, p. 20; Ingvar, 1979; Svoboda et al., 2006; Tse, 2013; 2018, p. 189). Significantly, many of these regions overlap with those of interest for the 'Dual Network theory'—system one and system two reasoning.

Recent research on meditation practitioners has shown increased nerve fibre and functional connectivity within the DMN, suggesting changes through activity dependence and a potential link between introspection and the neural processes necessary for top-down control (Shen et al., 2020). These findings support the notion that self-organization extends to the DMN and implicates crucial aspects of our mental functioning, explaining research that suggests this network is also closely tied to important aspects of moral decision-making, such as the theory of mind (Buckner et al., 2008, p. 23; Pujol et al., 2012, p. 918).

The full intricacy of the Default Mode Network (DMN) and its interplay with other brain networks remains an enigma, encompassing fragments of internal experience such as daydreams, musings, vibrant imagery, and meandering monologues (Buckner et al., 2008; Klinger, 1971, p. 347). This, as a reflection of a mysterious and distinctively human experience, signifies much of our cognitive existence—a form of inward processing independent of influences from the environment that remarkably, as we will see, allows us to extend our minds through time and space.

Changing the Mind—Not Two, but Three Ways

Studies such as self-organization and its foundational brain regions, including the DMN, highlight the necessity for further exploration to unravel its intricacies and understand vital facets of our rationality, as well as elucidations of how mental alterations occur. It seems untenable to

assert, as Levy does that there are "two basic ways to go about changing someone's mind" (Levy, 2007, p. 69). I propose a third method, which I term 'internal mentation,' and I will account for this in addressing parity and punishment equivalence in discussions that follow.

As with many of the matters discussed here, I return to disclaim the vast complexity of the brain and our limited understanding. Our comprehension remains provisional, and many processes continue to confound us. Perhaps the answer we seek—if they are to be found—falls squarely within the realm of cognitive and affective neuroscience. But I pose the challenge motivated by Klinger, who says of the unique introspective features of our mental life, "their very humanness lends them great intrinsic interest; but beyond that, indeed so prominent a set of activities cannot be *functionless*" (Buckner et al., 2008; Klinger, 1971, p. 347). Again, I offer not a scientific problem but a conceptual one, and I think it should motivate discussions about the parity principle, the scope of our rationality, and, in some more general sense, the potential promise and perils of neurotechnologies.

Mental Time Travel: Extending Rationality Across Temporal Horizons

This dissertation opened with a vivid illustration of Andy Dufresne's experience in solitary confinement (Darabont & King, 1994a; Darabont & King, 1996; King, 2010). Notably, his assertion that he had internalized Mozart's The Marriage of Figaro: Duetto-Sul Aria, which he played with the aim to 'free' his fellow inmates from the harsh realities of Shawshank prison. This mental strategy facilitated his endurance of the crushing consequences, making it the 'easiest time I ever did.' That by some strange mental event, the echoes of a young composer two centuries prior reached him and allowed him to turn his attention away from a dark place that we have otherwise seen is profoundly destructive to the mind and brain.

Indeed, while this is a fictitious illustration, it underscores a familiar phenomenon that indubitably plays a significant, frequently poignant, and inspiring role in our mental lives. There is a facet that remains inexplicably elusive not merely scientifically but also in philosophical discussions about the 'mind.' These discussions centre on difficulties explaining seemingly irreducible features of our first-person conscious experience that encompass both subjective and qualitative properties. It also includes 'intentionality,' the directedness of mental states towards something—which may extend not only to a different location but also a different place and *time* (Chalmers, 2007; Crane, 2015; Dennett, 1990; Nagel, 1974; Searle, 1983).

Crossing into the realm of science, 'mental time travel' is hailed as one of mankind's most remarkable feats (Yue et al., 2021). It intertwines with the declarative memory system and the DMN, introspection, and 'spontaneous internal mentation,' all vital to rationality and moral reasoning (Buckner et al., 2008, p. 20; Buckner & Carroll, 2007; Glannon, 2019b, p. 5; Ingvar, 1979; Svoboda et al., 2006; Veselis, 2017)

These mental states, Glannon describes, possess both a 'pastness' and a 'future-oriented aspect' (Glannon, 2019b, p. 5). They facilitate the conceptualization of a self that spans across time, distinct as past, present, and future selves, yet forming a unified identity (Siegel, 2012, p. 14; Yue et al., 2021). This preserves a sense of continuity and coherence and our experience of existing through time (Arstila, 2014; Caldas & Bertero, 2012; Siegel, 2012, p. 14; Yue et al., 2021). Our grasp of the temporal aspects of mental states invites parallels with the notions of a continuous, temporal existence—and the idea of the 'merging of horizons' which intertwines past, present, and future in our lived experience (Heidegger, 1962 [1927]).

But most significantly, embarking on a remarkable journey through time, 'autonoetic consciousness' grants us a first-person perspective to vividly revisit and re-experience past events, complete with the accompanying emotions. It also allows us to venture into the uncharted territory of the future, enabling us to envision and imagine potential scenarios yet to unfold (Michaelian, 2016; Michaelian et al., 2016; Zawadzki & Adamczyk, 2021, p. 8). And in undertaking this process, it seems that we *feel* something, and at least sometimes, it moves us, and something 'changes.'

Memories, particularly core ones, evoke strong emotions and shape our self-narratives (McAdams, 2013; McAdams & McLean, 2013; Zawadzki & Adamczyk, 2021, pp. 11-12). One common and powerful narrative depicts a 'redemption sequence,' transforming negative experiences into positive self-perceptions. Some even posit our propensity for storytelling evolved with higher executive functions in early humans, as suggested by discovered forms of symbolism (Foster, 2015, p. 26).

These residues of past experiences influence our present and future actions, contributing to the continuous story of 'self.' This self is one we weave through recursive, emergent processes of introspection, perception, and action that extend into our physical and social environment and through the expanses of time. Certain formative memories can even shape our mental architecture, aiding us in decoding the chaotic informational flux—'mental representation' that

'shapes how we think and act in the present and future' (Glannon, 2019b, p. 5). They change the mind and brain, and this should be of interest to the Parity Principle.

If we factor in temporal dynamics when assessing 'conscious control' and 'rationality,' it becomes clear that accomplishments like Hof's are not the result of a single moment but are rather the outcome of extensive training over time. This also applies to practitioners of meditation. Moreover, Hof's extraordinary feats are within our own potential. Starting with simple examples, studies suggest that embracing a nutritious diet, making healthier lifestyle choices, or even indulging in classical music can contribute to lowering blood pressure—a form of 'top-down control' over otherwise largely autonomous functions (Jenkins, 2020; Juneau, 2018; Moghadisian & Eskin, 2012; Trappe, 2010).

This extends to crucial decisions in our lives, where we consider our past experiences and our current circumstances and consciously steer our course through a sequence of intentional choices—some of which, I believe, are to steer clear of circumstances where we are prone to distraction and manipulation. In my view, we cannot comprehend what is happening in our brains at a given instant without taking into consideration the influence of such a chain of 'conscious' decisions. These decisions can themselves be designed to avert external interference.

Similarly, an immediate mirroring of effects between a neurointervention and a psychological method might suggest parity in their impact on or engagement with rational capacities. But our rationality and decision-making processes, woven from past experiences and directed toward future goals, must not be compartmentalized into single instances of neural or behavioural interventions. These processes unfold continuously through time, shaped by the interplay between past experiences and future aspirations.

Interestingly, an extensive focus on present decisions in assessing the effects of neurointerventions more generally, I think, reflects a form of Temporal Effect Bias (TEB) or 'time discounting' (Berns et al., 2007; Frederick et al., 2002). TEB sometimes misleads us to focus on immediate effects while undervaluing the long-term consequences of interventions. The extent to which this is the case, I set aside for another day. The point is simply that in neuroethics theorizing, we should not consider 'conscious control' or 'rationality' in this sense—nor the potential risks neurointerventions may present.

Embedded Rationality

A theme up to this point is how the mind is not isolated but a relational and social organ in constant flux and defined as a dynamic process. It maintains recursive relations between the brain and the physical and social environment, receiving neural signals from other brains. These interactions occur through patterned, recursive, and emergent self-organizing processes, where the law of complex systems applies to a single mind or two minds acting as a unified system (Fuchs, 2004, 2008, 2011; Fuchs & Schlimme, 2009; Glannon, 2009; Kornfield, 2009; Siegel, 2012, p. 12; Whitehead, 1929/1978). We discussed this by considering the four functional domains across which the mind operates—embodied, enacted, extended, and embedded (Ross et al., 2007; Varela et al., 2017).[102]

Because the mind supervenes on the brain, there is "a continuous transaction between current states of the brain, body, and the environment" (Clark, 2008a; Pouw et al., 2014, p. 53). Changes to the environment can change the brain, and this includes changes not only to the immediate physical but also to the social and *normative* environment, which I define using naturalistic descriptions of discernible normative properties. These properties are grounded in wider views on reasons, values, and relationship structures, reflecting "the ways in which both reasons and values depend on us: on our nature as rational, reason-responsive creatures, and on the social practices which we create or find ourselves embedded in" (Andrews, 2020; Heuer, 2022, p. 17; Raz & Heuer, 2022).

I propose that the normative environment—conveyed through cultural artifacts such as written language, religious symbols, and art—is deeply integrated into our cognitive processes, shaping our comprehension of the world (Donald, 1991; Whitehouse, 2004). These artifacts enhance communication, facilitate complex thought, aid problem-solving, bolster memory retention, shape moral reasoning and emotional responses, and embody 'distributed cognition,' all influencing brain functions (Clark, 2008b; Hollan et al., 2000; Hutchins, 1995). This is supported by neuroimaging studies identifying measurable changes occasioned by factors including education, cultural norms, religious beliefs, and language (Ansari, 2012; Baggio, 2018 Ch 8; Focquaert & Schermer, 2015, p. 141; Han et al., 2008; Skyrms, 2010).

In analyzing the temporal aspects of mental states and the journey of information from the brain via intertwined domains, including the normative environment and cultural mediums, it's evident that these cognitive operations could be subject to emergent principles such as self-

organization and nonlinearity. This control might manifest as patterns of distribution across time, capturing an extensive 'community of minds', both past and present, each contributing traits such as self-organization through internal mentation.[103] And this myriad of processes sculpts and changes the brain and the mind in powerful ways.

The extended mind thesis (EMT) and the parity principle underscore the importance of recognizing the mind's extension beyond the brain, incorporating external objects and being co-realized by external carriers (Clark & Chalmers, 1998; Heinrichs, 2018; Levy, 2020). Considering the profound influence of changes in the physical and social environment on shaping our minds, ethical considerations are warranted in relation to parity and the normative implications of brain interventions.

Particularly significant is the concept of "rational capacities," which forms the core of the debate concerning the parity principle and normative equivalence. If we acknowledge the mind's extension into the social and normative environment, it becomes essential to recognize that our mental capacities, including rationality, also extend into that environment, with all the implications that follow.

I recognize the parallel concerns between the extended mind thesis (EMT) and the notion of extended rationality. The risk of adopting an overly liberal understanding of what constitutes 'rational capacities' in a functionalist manner,[104] where function is prioritized over physical aspects, poses similar concerns as those regarding the extension of the mind. Although we could see rationality as scaffolded throughout time like some form of 'collective mind' or 'collective rationality, this does not seem to be sufficient—although technologies like brain-to-brain interfaces and AI might raise interesting hypothetical fodder for such discussions. The determination of ontological boundaries for 'rational capacities' is likely to be a topic of substantial debate in both theoretical and empirical discussions for which the parity principle would need to address (Adams & Aizawa, 2001; Heersmink, 2016; Heinrichs, 2018).

But as Levy correctly identifies, drawing on externalist ethics, the fact remains our agency extends into or becomes embedded within the world. (Levy, 2017; Levy, 2019; 2020, pp. 45-46). And this is something we must account for in considering a broader range of implications of using neurotechnologies to circumvent rational agency.

The Inner Sphere, Our Path Through Time, and Our Place in the World

In wrapping up this discussion, I stress the intricate interplay between our mental capacities and potent environmental influences, which underlines not only the boundaries but the vast potential of our rationality. To effectively navigate the nuanced dilemmas posed by direct and indirect manipulations of the mind, we must critically scrutinize the foundational tenets of our mental existence and presuppositions about rational agency.

The agent-centred view, which holds prominence in action theory, often asserts—albeit in various forms, that human rationality can be understood in terms of mental states such as beliefs, intentions, and desires and the critical role of 'reasons' in guiding decision-making, shaping rational action, and eliciting attributions of moral and criminal responsibility (Fischer, 1998; Fischer & Ravizza, 1998; Frankfurt, 1988a; Greene & Cohen, 2004; Kalis et al., 2008; Mele, 1992, 1995; Morse, 2011; Morse, 2000, 2004; Parfit, 1984). However, we should remain open to the possibility that such accounts do not correctly, fully or accurately depict the scope or key features of human rationality—a sentiment shared by numerous scholars (Brown, 2015, p. 46; Hardcastle, 2020, p. 157; Lewis, 1970, 1972; Lewis, 1966; Northoff & Wagner, 2018, p. 345; Whitehead, 1929/1978).

Intriguing insights spring from distinctive scientific phenomena like self-organization, temporality, and our cognition's complex social dynamics. Unravelling these phenomena could help us gain a deeper comprehension of inherent complexities and probe the potential moral asymmetry within the realm of mental alteration.

Our foray into the Default Mode Network (DMN), mental time travel, and our entrenchment in social and normative contexts discloses striking facets of our distinctly human mental existence. They provoke crucial questions about the myriad ways our minds are capable of self-organization, expand through space and time, and are embedded in a larger social and normative context.

Reflecting on internal mentation, our brain exhibits a remarkable ability to self-organize and rebuff external stimuli, even in moments devoid of any interventions. The mental shifts in meditation practitioners, for instance, can't be attributed solely to environmental stimuli, indicating an inherent capacity for resistance. This inner space, so critical for rationality, defies simple explanation by 'direct or indirect' interventions.

What allows us to journey through time and space within our minds, pondering our past, present, and future selves? This extraordinary capacity cannot be fully explained without considering how it evolves over time, shaped by a series of decisions that forge our rational abilities. Remarkable examples like Hof's feats—one accomplished not just in an isolated moment but over a series of conscious decisions— highlight the power of 'conscious control' and the similar power we have in our own lives through utilizing rational decisions in moulding our minds, relationships, and life trajectories.

Importantly, our mental processes are not solitary enterprises. They weave into a broader social and normative tapestry. Our minds, shaped through introspection and shared experiences, exist within a dynamic network of interlinked minds that mould history and chart future courses.

Given the complex, extended and embedded system of the mind, alongside emergent properties, such as self-organization and bidirectional causation, it is worth exploring the extent to which all of these interact. When intervening directly in the brain, we must acknowledge that this bypasses our 'rational capacities.' If rational capacities are seen as a set of mental abilities entirely subject to sensory data and environmental forces, perhaps there is nothing inherently objectionable about such interventions.

But if we conceive of 'conscious control' and 'rationality' in a broader sense, we need to delve deeper into ethical considerations. Might this intrude in an 'inner sphere' that, at least sometimes, remains impervious or even resists environmental influences? How might such interventions affect our perception of self over time? How could they reshape our relationships within the social and normative community? And for discussions in the next chapter, what sorts of moral or ethical considerations ought to protect persons in the face of such threats—even those guilty of wrongdoing—and how might such considerations find footing in the realm and discourse of rights?

Let us be clear: indirect interventions and environmental influences can profoundly distort our mental lives, disrupting our ability to represent the world coherently and self-regulate effectively. They can distract us, impede learning from our past, and undermine our focus on the future. However, understanding rationality and the scope of human freedom in this light does not mandate such negative outcomes. Instead, it suggests a potential for transformation, leveraging these very capacities to recognize their limitations and, perhaps, surmount them.

Concluding Reflections: Embracing Rationality by Acknowledging Irrationality

Let us circle back to our initial inquiry: Do direct brain and mind interventions fundamentally diverge from indirect methods in altering rational faculties? The contentious debate on this matter necessitates a comprehensive evaluation of the ethical divergence between these approaches, considering their mind-altering effects and potential threats to human liberty.

So far, I have kindled this discourse through vivid examples—be it music echoing through prison yards, astonishing cold endurance feats, meditation practitioners, cybernetic cockroaches, or evocative allusions to unbounded freedom or a dystopian post-human future. However, the essence lies not in these illustrations but in the philosophical conundrums they stimulate—conundrums that are not just unusual but also enthralling and, I think, more intriguing. The intention here is not to offer a resolution but to foreground key puzzles that I deem significant for future theorization and underscore the necessity to address them.

Moving forward, more in-depth theorizing and a recalibration of our focus on these critical facets are required to broaden our comprehension of rationality, interventions, and the extensive ethical implications of forced state intervention in criminal offenders' brains and minds. This might shed light on moral or normative reasons for providing them with a broader protective ambit based on credible normative theories possibly articulated in terms of moral or legal rights—an issue we turn to in the next chapter.

But I conclude with the following remarks. Beneath the surface of our conscious awareness, we are ensnared by a potent undercurrent, emitting mists of illusion, fears, and deceptive apparitions, all the while encompassed by powerful distorting and subversive forces that beleaguer us at every turn—perhaps at this very moment. While we usually perceive ourselves as free and rational beings, we remain incognizant of our own rationality's profound limitations. However, this does not preclude us from gaining self-knowledge and learning about these shortcomings.

It is true we are, like everything else, 'subjects of the world.' Acknowledging the severe limitations of our rationality need not engender despair. Just as Darwin correctly foresaw his contemporaries would grapple with acceptance of the radical concept of evolution, we, too, must confront our rationality's profound limitations.

However, as Paul Davies notes, building on Darwin's rhetoric, embracing these constraints can engender a 'love for nature and life' and an appreciation for the strong emotions that render us 'strangers to ourselves,' which simultaneously present us with a 'gift for living.' This perspective can enable us to perceive ourselves as we genuinely are (Darwin, 1860/2010; Davies, 2009, 2020; Panksepp, 2004; Wilson, 2002).

But as 'subjects of the world,' we are also components of a universe that, for reasons unbeknownst to us, have attained self-consciousness and self-awareness (Nagel, 2012). Amid a biological substrate, whose configurations over time even surpass those of the particles in the universe from which it springs, it is nothing short of extraordinary that this same universe has not only become conscious and aware of itself but conscious and aware of the limitations of limits the limits of its own consciousness and rationality.

Acknowledging and examining one's own irrationality can be considered a form of rationality in itself (Ariely, 2010; Tavris & Aronson, 2007). Perhaps we owe credit to the research we have considered to this point that may allow us to do so—theories and research that reflect cultural artifacts in our embedded environment. Research suggests that our beliefs about our own agency and rationality can significantly influence our actions and even our brain processes. For instance, participants who were exposed to deterministic texts cheated more on a subsequent math test than those who read neutral texts. When we perceive our actions as predetermined, we may feel less responsible for our behaviour, leading to changes in how we act (Vohs & Schooler, 2008). I suggest the same might hold for acknowledging and learning about the limits of our rationality and striving to address them as we navigate and shape our lives.

Earlier, I suggested a potential path to transcend our limitations. This resides, I believe, in our ability during serene moments of introspection to accept and understand the limits of our own rationality—not solely within our own minds but as part of a collective endeavour to conceive of the world as we might reasonably aspire it to be. In so doing, our minds are altered, a realization that carries implications not only for our theorizing but also for the biological transformation of our brains, minds, and the expansive social and normative environment in which they exist and will continue to persist—perhaps long after we do not.

4

The Equivalence Claim and Mental Freedom

Freedom to Fall and the God Machine

Introduction

As we begin this final chapter, we navigate a philosophical landscape steeped in history, encapsulated in the cultural artefacts reflecting our collective psyche—the scope and limits of human freedom. Across cultures and epochs, thinkers have grappled with the delicate equilibrium between an aspiration for freedom unfettered and the hidden demons of human nature—the rage, the lust, and the lawlessness "let off the chain" (Dostoyevsky, [1880] 2003). The discourse reminds us that while "Man is born free, and everywhere he is in chains" (Rousseau, [1762] 1964)—we are not free in the ways we generally suppose. But it also reminds us that freedom calls for a deep sense of responsibility, an echo of Nietzsche's perspective (Nietzsche, [1844-1900] 2006), and in its fullest sense, this may lie in a way of life that respects and enhances the freedom of others.

As we embark on this exploration and near the end of our journey amidst the evolving landscape of neuroscience and neurotechnologies, we delve into the profound dimensions of freedom, responsibility, and societal order. Our journey brings into focus the nature of punishment and its potential transformations in an era of rapid advancements. This requires us to carefully consider whether the brain, our cornerstone of identity and rationality, symbolizes a realm of personal freedom demanding a unique form of protection. Whether neurointerventions may pose distinct threats—especially for society's most marginalized, like prisoners or norm violators—and intrude on a realm that ought to remain beyond the purview of the state—"something inside that they can't get to, that they can't touch" (Darabont & King, 1994a). The central point of interest that I consider is 'human freedom'—and perhaps most fittingly, a form of 'mental freedom'.

Recap—Punishment Equivalence and Asymmetry

This dissertation delves into a provocative question: if we incarcerate criminals without choice, why not mandate brain interventions? We unearth intricate issues that surface under scrutiny by exploring punishment equivalence's key aspects—the Punishment Claim, In-Principle Constraint, and Parity Principle.

In chapters 1 and 2, in deconstructing the Punishment Claim and In-Principle Constraint, I identified a host of contingent empirical concerns around the failures of current punishment practices, a history of abuse of prisoners, our primitive neurobiology, the magnitude of assumptions required for claims of safety and efficacy—and the catastrophic risks of a rush to forcibly administer neurointerventions in our criminal justice practices.

Chapter 3 considers issues that I see as posing conceptual challenges to the 'parity principle.' I identified concerns neurointerventions are 'arational'—they have the ability to 'circumvent rational capacities' in a distinct way conventional forms of punishment do not—raising concerns they are 'freedom subversive' (Focquaert & Schermer, 2015, p. 141; Harris, 2014b, p. 372). However, I also contended that the parity principle neglects vital aspects of human rationality, suggesting we might need to broaden our perception of human freedom and potential risks posed by neurointerventions.

Our exploration culminates at the Equivalence Claim, suggesting neurointerventions could supplant conventional forms of punishment—like incarceration—because their effects are equivalent or preferable—being 'less wrong' or 'more beneficial. Contrarily, the Asymmetry Claim spurs us to question the real comparability of these punitive forms—they present a distinct wrong.

But to appraise the 'less wrong' or 'more beneficial,' we need to say more. We must, I think, adopt a specific comparative metric. But this also requires us to draw important distinctions that are not always clear or expressed in the present ethical discourse on neurointerventions and punishment.

First, recall the difference between *ideal* and *non-ideal theory*. Ideal theory employs reasoning based on abstract principles and perfect scenarios, ignoring practical constraints and real-world complexities. Conversely, non-ideal theory considers practical limitations and empirical facts in ethical analysis.

This traces a related distinction I have drawn between moral asymmetry and normative asymmetry. Moral asymmetry focuses on broader first-order moral considerations about deontic constraints, fairness and rights—which generate moral reasons that inform analysis. Normative asymmetry is broader. It includes these considerations but also practical aspects such as safety, efficacy, and legal implications. The latter generates both moral and *practical* or *prudential* reasons that may support or undermine claims of equivalence.

Attentive to these distinctions, theorizing about whether there is something 'wrong' about neurointerventions, compared to conventional punishment, may, and often does, fall in various domains—normative theory, political theory, and applied fields like public policy, bioethics, and law. Each domain offers distinct analytical tools and conceptual frameworks, accounting for any range of ideal and non-ideal considerations and corresponding moral and normative judgements. Different results may follow. The problem is the boundaries between them are not always clear, complicating equivalence and asymmetry assessments.

In this chapter, I trace two lines of analysis. First, I hope to identify ground analysis in a way that illustrates the inherent complexity and challenges that arise when these distinctions are not clear and expressed. The second is to pose a challenge to equivalence claims by identifying what I see to be the most compelling source of moral and normative asymmetry—the unique threats neurointerventions pose to 'mental freedom'.

Before we begin, recall arguments for punishment equivalence do not necessarily assert the ethical or moral permissibility of neurointerventions. Instead, they challenge objections that might hinder their mandatory implementation in the criminal justice system. In so doing, I have described them as 'permissive'—they lower barriers that might otherwise challenge our intuitions there is something inherently wrong about them. In at least some cases, they claim to have displaced or cast down on one of the 'stronger' or 'most compelling' objections.

The analysis here adopts a similar approach. Building on previous chapters, I argue the unique threats neurointerventions pose to 'mental freedom' presents the stronger and most compelling argument against neurointerventions. The goal is not to explore every line of analysis. Rather, it is to identify what I see to be pressing areas that require attention. At least at present, equivalency arguments are not successful in assuaging concerns that such threats present a serious risk of asymmetry, which permeates various domains of analysis.

At the very least, I hope to illustrate the need for great caution. Before we venture to assert conclusive or ambitious claims about sanctioning large-scale social projects that use neurointerventions to address crime, or before we consider sweeping changes to our punishment practices, we must clearly identify and address looming sources of moral and normative asymmetry based on threats to mental freedom. Such a task calls for accountable, thorough, and transparent theorizing—the stakes are too high for anything less.[105]

The Freedom Objection and the Freedom to Fall

A fundamental and widespread critique of neurointerventions, sometimes addressed at the highest echelons of ideal and metaphysical discussion, is the 'freedom objection'. This objection claims asymmetry between conventional punishment and neurointerventions, bringing to light a type of freedom crucial to human virtue and development.

John Harris, a vocal proponent of the freedom objection, delineates a form of freedom that imparts valuable perspectives into our dialogues (Aristotle, 2009; Harris, 2011, p. 104; 2013b; 2014b, p. 75; John Harris, 2016). He designates this as the 'freedom to fall'—a notion borrowed from Milton's Paradise Lost, which reflects Judeo-Christian theology and the fall of humankind from Eden, a consequence of Adam and Eve's exercise of free will and their acquisition of knowledge (Harris, 2014b, p. 373; Milton, [1667] 2014). In the biblical story, Adam and Eve decided to act against God's will and thus exercised their freedom to err. Harris describes this form of freedom as the "freedom to decide whether or not to fall for reasons, which have to do with what is best 'all things considered'" (Harris, 2014b, p. 373; Milton, [1667] 2014).

The potent allure of the 'freedom to fall,' is evident in Milton's conception of one who might consciously prefer to reign in Hell for the sake of freedom rather than serving in Heaven—"Here we may reign secure, and in my choice to reign is worth ambition though in Hell: Better to reign in Hell, than serve in Heaven" (Milton, [1667] 2014 Book I, lines 258-263). Beyond such conceptions, it also traces themes from Milton's classic work about how such freedom is necessary for 'genuine allegiance, faith, and love' to be manifested through voluntary decisions rather than coerced actions (Anderson, 2010; Milton, 1667; Milton & Hughes, 1957, p. 837; Savage, 1977, p. 286). These concepts, in philosophical nomenclature, trace themes of 'virtue cultivation,' which suggest 'falling' is an essential aspect of our moral growth, allowing us to

learn from mistakes and develop as individuals. And in the larger debate, frame discussions about how neurointerventions might limit or restrict this ability.

We will begin by considering the 'freedom to fall' as a metaphysical concept, as some have formulated it (Pugh, 2019; Pugh, 2020). This is a useful starting point. But as we will see, 'freedom' is far from a singular, monolithic concept; rather, it is a multifaceted notion, spanning various dimensions and carrying different implications across different contexts.

Freedom of Will—Free Choice and Free Action

Discussions about human freedom at the metaphysical level often revolve around the free will debate—whether or not we possess 'free will.' The 'will' in this context is a philosophical construct associated with agent-centred accounts of rational capacities, embodying beliefs, intentions, desires, and the ability to respond to diverse circumstances (Dennett, 2015; Descartes, 1641/1996; Dworkin, 1988a; Fischer, 1998; Fischer, 1999; Fischer, 2006a; Fischer & Ravizza, 2000; Frankfurt, 1971; citing Frankfurt, 1988a, 1988b; Glannon, 2018a, p. 319; McKenna, 2000, p. 97; Quante, 2011; Spence, 2009). Generally, to say the will 'is free' is to say persons have the ability both to choose and act according to their will (Hume & Millican, 1748/2007). Again, I set aside the issue of whether this conception of 'freedom' accords with scientific views of human rationality based on our discussion in previous chapters (Brown, 2015, p. 46; Hardcastle, 2020, p. 157; Lewis, 1970, 1972; Lewis, 1966; Northoff & Wagner, 2018, p. 345; Whitehead, 1929/1978).

Notwithstanding, I think, in general, notions of 'freedom' inevitably grapple with determinism—the proposition that all events, including human actions, are preordained by preceding events and natural laws (Pereboom, 2014, 2018; Smilansky, 2000; Strawson, 1994, 2010). If determinism holds, then our thoughts, actions, and behaviours are ultimately influenced by the universe's laws and past events. According to this view, our perceived freedom of action is but an illusion despite our conscious experience.

Yet, our perception of morality and our reactions to others' behaviours are deeply intertwined with our conception of freedom and the 'will.' For example, it is often thought moral accountability presupposes alternative possibilities; an individual 'could have done otherwise' (Frankfurt, 1969).[106] It also requires 'regulative control'— "the sort of control that involves

genuine metaphysical access to alternative possibilities" (Fischer, 2006b)—being freedom of choice *and* action.

This is significant because prevailing normative frameworks do not typically assign responsibility for actions over which an individual has no control. Standard views of moral responsibility, for example, posit that individuals may be morally responsible for their actions or otherwise properly subject to desert or 'reactive attitudes' such as praise, respect, indignation, and forgiveness— insofar as they knowingly and deliberately chose a given path in the face of viable alternatives (Olsaretti, 2003; H. Simmons, 2010; Smart, 1961; Strawson, 1974; Vihvelin, 2022). In fact, it is also possible determinism might render certain views of punishment, such as retributivism, misguided (Pereboom, 2001, 2014, 2020).

Some accept the truth of determinism and that we are not free in the ways we generally suppose. Others reject determinism. One compelling rejection draws on insights from quantum physics, suggesting the universe, and by extension, human actions, might not be fully determined (Fischer, 1999; Fischer, 2006a; Fischer & Ravizza, 2000; Tse, 2013, 2018). However, the question arises: does this indeterminacy merely render our choices products of 'luck'? Or does it still allow for a degree of control necessary for freedom itself (Haji, 2016; Levy, 2011; Levy & McKenna, 2009; Mele, 2006)?

A different philosophical perspective—termed compatibilism—proposes that the truth of determinism need not undermine our understanding of free will. Certain theories within this school suggest that determinism and freedom might be joined if we distinguish between freedom of choice and freedom of action. Some compatibilists dispute the common intuition that moral responsibility necessitates the capacity to act differently. Instead, they argue that it might be enough to have the freedom to choose, even if one lacks the capability to enact these choices (Dennett, 2015; Frankfurt, 1969, 2018; Haji, 2009; Robb, 2022; Widerker & McKenna, 2003). [107]

While we are only scratching the surface of the expansive free will debate here, honing in on the perceptions of 'freedom' from a high-level perspective, our aim is to identify the central questions which are present in the literature. These include questions such as: are we genuinely free? What conditions are needed to fulfil this freedom? What sort of restrictions undermine it? How does this freedom shape our moral responsibility? From this angle, intertwined in each question is the 'freedom to fall.'

Neurointerventions and the Freedom to Fall

Harris introduces the 'freedom to fall' and the 'freedom objection' to identify what he sees to be the most significant threat of neurointerventions—threats to human freedom. What is significant is the "freedom to fall" encompasses not only the *choice* to do wrong but also the ability to *do* wrong – to be free to *fall*. The following section will make manifest the importance of this kind of freedom, including both aspects of its nature, arguing that the ability to err is something that has deeply valuable aspects which warrant fervent protection.

Harris' 'freedom to fall' is motivated by concerns about neurointerventions—in his case, as part of larger social issues about mandatory moral bioenhancement. His 'freedom objection' reflects a critique I mentioned in the previous chapter, that neurointerventions are 'passive,' 'a-rational,' can bypass rational faculties, and hence are 'freedom subversive' (Harris, 2014b, p. 372). He portrays these interventions as acting "directly on the mainsprings of action, on emotions or other dispositions," thereby circumventing "what they perceive as a dangerously paralyzing or dilatory process that might somehow get between an impulse and the moral action it impels" (Harris, 2012, p. 294).

For our purposes in this chapter, Harris' freedom to fall provides a valuable entry point for exploring a potential moral asymmetry between neurointerventions and conventional forms of punishment. The freedom to fall and its associations with 'mental freedom' and the 'freedom to do wrong' informs salient issues about the potential threats posed by neurointerventions in the domain of criminal punishment.

The Freedom to Fall—Morality and Metaphysical Conceptions

If neurointerventions bypass our rational faculties, as I propose, and infringe upon human freedom in a distinct way, the 'freedom to fall' could be seen to offer the most fertile grounds of inquiry into identifying a *moral asymmetry* between neurointerventions and conventional forms of punishment. In addressing 'freedom,' let us begin at the apex of ideal theorizing and metaphysical discourse. If we accept the premise that neurointerventions infringe upon the 'freedom to fall,' which is interpreted as the metaphysical liberty to make choices and perform actions, including regulative control (Fischer, 2006b) and the ability to act wrongly (Harris, 2014b, p. 373; Milton, [1667] 2014), what implications might this have? I think there are three places we might look.

The first concern arises from the concept of 'rationality' I endorsed in the last chapter. Very generally, that the 'will' and our 'rational capacities' comprise an inner sphere, typified by 'internal mentation'—at least *in principle* uniquely resilient to conventional intervention—is expressed through mental states with temporal aspects, persisting and extending through space and time, and our agency is scaffolded and embedded into a broader social and normative discourse. In turn, I think that human freedom involves, or at least implicates, various facets of our mental life, interpersonal relationships, and even structures of social and normative relationships. In this sense, I have hinted that diminishing regulative control through neurointerventions could have implications that resonate across these various domains—perhaps in a more widespread manner than might generally be supposed.

If we accept this, I think there are implications for discussions of 'metaphysical freedom' at the highest level of analysis. First and foremost is the deep moral significance of metaphysical freedom. Our emotional responses and notions of free will are intertwined, as they revolve around the belief that we have control over our actions and could have done otherwise (Kane, 1998; Olsaretti, 2003; Pereboom, 2014; H. Simmons, 2010; Smart, 1961; Strawson, 1974; Vihvelin, 2022; Wolf, 1990). If neurointerventions interfere with rationality, freedom of choice, and action—the enacted nature of mental states—they risk disrupting our moral and emotional landscapes and likely the equilibrium of our social and normative relations. This may lead not only to a disruption of direct individual attributions of moral praise, blame, and reactive attitudes but also to the larger discourse in which those judgements are grounded.

The second concern is that having the ability to make choices and shape one's path, even at a metaphysical level, is closely tied to a life of significance and purpose. Kant believed there is something valuable about a self-determined life as an end in itself (Ivison, 2007, p. 97; Kant, 1999 (1781); Lee, 2012, p. 338; Stolzenberg, 2008; Varga, 2015, p. 74). Neurointerventions that limit this freedom may impact the existential dimensions of human life. The intrinsic value of metaphysical freedom lies in the ability to author one's life and have authentic control over choices and actions (King, 2014; Olsaretti, 2003; H. Simmons, 2010; Smart, 1961; Smilansky, 1996; Strawson, 1974; Vihvelin, 2022). Diminishing this control through neurointerventions, those expressed through serene moments of introspection where we accept the limits of our rationality and manifest in our lives and as part of a collective endeavour, has significant

implications. It may be seen to diminish something inherently valuable and meaningful at an individual level.

A response might arise which points out that conventional forms of punishment similarly inhibit self-direction, autonomous action, and the ability to manifest a life of one's own making. However, the identification of internal mentation and the individual's ability to introspectively adapt to adverse conditions ensure separation. A person may be physically subjected to incarceration while retaining the ability to choose how to respond to and go forth in a manner of their own creation.

However, I acknowledge exploring the full scope and nature of these implications would benefit from more concerted analysis drawing on rich themes from the larger metaphysical debate on free will and moral responsibility—productively grounded, I think, in a scientific conception of the 'freedom' and the 'will.' And as we will see, these larger metaphysical issues have implications for moral and normative metrics traced across a broad spectrum and domains of inquiry—from principles such as humanness, respect, human dignity, political legitimacy, autonomy, personal identity, and authenticity.

But beyond this, I turn to what I see to be a third fruitful avenue of inquiry at the metaphysical level of analysis and one which appears to be of significant importance for Harris, Milton, and the 'freedom to fall'—*the cultivation of virtue*.

Freedom and Virtue

Harris argues the 'freedom to fall' holds a profound significance in our exploration of human freedom because it includes the 'freedom to do wrong.' The ability to make mistakes to deviate from the morally right path is an integral part of our journey toward virtue and moral growth (Harris, 2011, p. 104; 2013b; 2014b, p. 75; John Harris, 2016)

Delving into the realm of virtue theories, we move beyond the conventional focus on the rightness or wrongness of actions. Instead, the focus is on the profound significance of nurturing virtuous character traits and living a moral life that goes beyond mere decision-making. Virtue encompasses the essence of what it means to be human, shaping our character and guiding our judgment. Put another way, a 'moral life' involves 'much more than right and wrong decisions and actions,' and instead 'what human beings ought to be and the sort of life we ought to lead' (Adams, 2006, p. 3; Anscombe, 1958; Foot, 2002; Rüther & Heinrichs, 2019, p. 356).

Drawing inspiration from ancient philosophers like Aristotle, we understand that virtue is not a singular act but a practice—an ongoing, conscious effort entwined with our choices and actions. Building on the *Nicomachean Ethics*, this resonates with what Aristotle called "phronesis," or practical wisdom, which allows us to navigate the ethical implications of our actions with precision. Cultivating this wisdom leads to a state of "eudaimonia," a symphony of harmonious and ethical living (Aristotle, 2009).

What is fascinating is such notions of cultivating virtue from a distant Socratic era share parallels with the contemporary scientific perspectives we have explored. The dynamic, flexible, and adaptive capacities of the brain and mind, encompassing integration, autobiographical recall, and future envisioning, contribute to a cohesive representation of the world and enable self-regulation. These capacities are intricately linked to moral judgment, evident in specific brain regions and manifested throughout the lifespan of individuals across various domains in pervasive ways through activity dependence (Friston, 2010; Glannon, 2020, p. 90; Siegel, 2012, p. 9; Sporns, 2010; Tononi & Sporns, 2003). These features also highlight the crucial aspects we've identified as potential differentiating effects of neurointerventions.

But in this context, the freedom to do wrong takes on a deeper meaning. It is not about endorsing immoral actions but recognizing the necessity of the freedom to err. Without the ability to choose wrongly, we lack the essential learning experiences required for moral growth and the ability to navigate the ethical dimensions of our lives effectively. Morality, in this view, is not a forced compliance but an internalization of values through free choice over and against the freedom to err. The freedom to fall, to err, and to learn from those errors becomes indispensable on our path to virtue (Snow, 2010; Swanton, 2003). Along the way, moral failures or 'falls' can be transformative, leading to personal growth, increased empathy, and a more profound understanding of morality.

What I also think is important is how the cultivation of virtue at an individual level bears implications at a social level. We have identified the important connection between free will and moral responsibility—our ascriptions of praise, blame, obligation, and reactive attitudes (Kane, 1998; Olsaretti, 2003; Pereboom, 2014; H. Simmons, 2010; Smart, 1961; Strawson, 1974; Vihvelin, 2022; Wolf, 1990). But alongside such judgements, what we see as 'virtue' or 'virtuous action' are defined through "our relationships to… other people. And it is only by reference to [such] virtues that we can understand what real freedom is" (MacIntyre, 1981).

Further, all form part of a larger social praxis that mediates and structures between persons. The moral respect we owe to others is premised on our acknowledgement of their status as rational beings, making choices about the courses their lives will take (Lippke, 1998), and on some accounts, this a necessary condition to ensure a stable *socious* (Bublitz, 2018, pp. 316-317).

In this sense, the 'freedom objection' and the 'freedom to fall' add to our understanding of the *normative environment* conducive to the development of virtue. Expanding upon this idea promotes the notion of an 'externalist ethics' that could help dispel cognitive distortions that excessively focus on the demarcations between individuals. A recognition of our shared human traits, including our errors, should also extend to those who transgress societal norms, such as criminal offenders, whom we may otherwise be quick to condemn. Recognizing this shared humaneness could enrich our understanding of ethics, our collective experience, and the diverse expressions of human freedom.

What is fascinating, I think, is if we accept notions of embedded cognition—as I have endorsed—our very moral ascriptions might be seen to be, at least in some respects, a product of a community of minds that could equally be recursively refined through embedded mental processes. Within this normative environment, we can come to appreciate an internal symphony of interaction and influence that is not purely metaphorical but, in fact, causally prevails over our social atmosphere. Thus, I think, alongside Harris, that a *society* valuing virtue must value the 'freedom to fall: 'freedom to decide whether or not to fall for reasons [of their own accord]' as it provides the necessary space for moral growth. (Harris, 2014b, p. 373; Milton, [1667] 2014).

What comes out of this, for the purpose of neurointerventions, at the very highest level of metaphysical and conceptual analysis? I would only make a few brief comments that trace themes in discussions that follow.

First, at the individual, normative level, neurointerventions that restrict the freedom to do wrong may hinder the process of moral growth and impede our ability to effectively navigate the ethical complexities of our lives. Moral failures and the subsequent learning that arises from them play a crucial role in personal development, fostering growth, empathy, and a more profound understanding of morality. By impeding this freedom, neurointerventions risk circumventing the vital and delicate aspects of human rationality and freedom that contribute to this process. For example, consider the powerful core memories and emotions that shape our self-narratives

through recursive patterns across time, such as the 'redemption sequence' (McAdams, 2013; McAdams & McLean, 2013; Zawadzki & Adamczyk, 2021, pp. 11-12).

Moreover, the impact of such limitations extends beyond the individual realm and permeates the social fabric. Society relies on the collective learning and development that arises from the freedom to make mistakes, acknowledge failures, and collectively grow from them. This traces themes about the importance of our collective ability to reproach in *ourselves* the faults we criticize in others (Laertius, 1853, pp. 256-259). Recall our discussions throughout this work about our *collective* failures—egregious human rights violations involving prisoners, overcriminalization expressed in tragic historical cases and, more recently, the human rights revolution precipitated by the abhorrent events of the Second World War. It is through the recognition of our shared human traits, including our errors, that we can collectively shape our ethical standards and foster a deeper understanding of morality. By constraining the freedom to do wrong, neurointerventions disrupt this social praxis, depriving society of the power to learn, develop, and evolve collectively.

And all the while, I think, tracing the theme I identified at the close of the last chapter, what is truly unique about our capacity as human beings is our ability to recognize the limits of our own rationality—to come to see ourselves for who we truly are. As I have posited, this very recognition presents, at least in theory, perhaps the most potent catalyst for personal transformation, societal evolution, enhanced empathy, and a deeper understanding of morality. The essence of freedom is to err, to falter, yet to learn and evolve through these experiences. This growth is not confined to our personal lives alone; it extends to our collective missteps within the broader moral landscape. It is through our shared failures that we evolve, improving ourselves and the society in which we exist. In this way, we may work toward cultivating *phronesis*—practical wisdom—and, however ambitious, the pursuit of eudaimonia—human flourishing—(Aristotle, 2009).

Even at the highest level of abstract theorizing, I think these considerations highlight the need for careful deliberation when it comes to the ethical boundaries of neurointerventions. Certainly, virtue theories are subject to objections, including the challenge of defining virtues, identifying criteria for cultivation, and how or even whether virtues can be cultivated or taught (Annas, 2003, 2011; Heinrichs & Stake, 2019; Statman, 1997; Swanton, 2003). There may be valid reasons to place certain limits on human actions. But we must be mindful, as we have

throughout this dissertation, of the harms of crime and the corresponding interest in preventing harm to others.

It is essential to recognize the value of the freedom to fall and the potential risks associated with its restriction—particularly to the cultivation of human virtue. Virtue theories continue to contribute significantly to our understanding of ethics and the development of moral character, and I think at the highest level of analysis, they also bear significant weight in discussions of neurointerventions.

The God Machine and the Bounds of Freedom

Transitioning from ideals of *eudaimonia* and the biblical notions of 'the fall', let us journey into a speculative realm to the year 2050 under the omnipresent surveillance of an all-knowing entity: the 'God Machine.' Born from the fertile intellectual ground of Persson and Savulescu's imaginations, using advanced optogenetics and genetically modified neurons containing 'nano-signalers,' this celestial sentinel scrutinizes humanity's mental processes with relentless vigilance, ready to step in when one teeters on the precipice of severe immorality, all without their even knowing. This God Machine hailed as offering a semblance of 'near-complete freedom,' is envisaged as a sort of protective moral overseer, curbing our actions only as we verge on serious transgression (Savulescu & Persson, 2012). As the authors state:

> It is, perhaps, this kind of world which objectors to moral enhancement like Harris fear. Human beings are no longer 'free to fall' or at least not free to fall big time. But it might be wondered what is so bad with such a world after all? Those who value and want to be free can be free, or at least as free as humans can ever be. And everyone is much better off for the absence of evil. (Savulescu & Persson, 2012, p. 413).

The God Machine presents a formidable challenge to the "freedom objection" and moral asymmetry claims within the realm of ideal theorizing. Moreover, it has stimulated intense discourse and critical examinations in scholarly circles, as evidenced by the fervent critiques and defences of Harris' freedom to fall put forth by various authors (Bublitz, 2015b; DeGrazia, 2014; Douglas, 2013; Harris, 2014a; J. Harris, 2016; Hauskeller, 2017; Persson & Savulescu, 2015; Pugh, 2019; Savulescu & Persson, 2012; Young, 2019). These discussions directly align

with the central focus of this dissertation. By drawing parallels with advanced speculative or fictional neurointerventions aimed at crime prevention, the concept of the God Machine sets a solid foundation for further explorations and debates which pivot on theories of 'freedom' and 'falling'—or doing wrong.

I see the God Machine as presenting two challenges. The first is whether we should value both the freedom to choose *and* act—or if merely choosing is enough. The second is whether, even if we accept the freedom to act is important, why we should value the freedom to commit seriously immoral acts at all.

Free Choice, Free Action, and Freedom

Harris' 'freedom to fall' suggests that for moral virtue, we need 'regulative control,' or the freedom to *act* on our diverse decisions. The God Machine, inspired by Frankfurt-style experiments and 'compatibilist' views of free will, disputes this. It suggests that moral responsibility may be sufficient if it is based only on the freedom of choice, not the need for 'regulative control.' The God Machine asks us to consider whether, even if we limit certain behaviours, it does not truly infringe on our freedom or moral virtue. Perhaps, merely having the choice to do wrong, even if the wrongdoing is prevented, is enough to uphold moral responsibility—or at least ground the sorts of reactive attitudes or capture salient moral features that are significant.

I do not believe that the aspect of the God Machine poses a significant challenge to the concept of 'freedom to fall.' The God Machine's focus on protecting a limited form of 'freedom' is insufficient to alleviate concerns about moral judgments and the development of moral virtues. It fails to address the importance of freedom of action, which is valued in theoretical, metaphysical, and moral domains.

Arguments Against Source Compatibilism

First, the God Machine builds on a 'source compatibilist' viewpoint, suggesting that free will and some moral assessments can exist with merely the freedom of choice, regardless if the choices are never realized. This situates the God Machine within intense philosophical debates and unresolved issues linked to source compatibilism. I believe the God Machine's stance on 'freedom' is built on a shaky, or at least highly contentious, philosophical ground, as it inherits criticisms that might be generally attracted by source compatibilism. For example, one critique

might be its failure to present sufficiently robust alternative possibilities, crucial for assigning moral responsibility—particularly given the God Machine's persistent endangerment of any such alternatives (Fischer, 1994, 2006b; Fischer & Ravizza, 1998, 2000; Pereboom, 2009; Vihvelin, 2022).

We cannot trace all of these lines of inquiry, but I simply note despite decades of debates and counterarguments since Frankfurt's initial thought experiment, a clear consensus has not been reached—leading some to describe the situation as a 'stalemate (Speak, 2005; Vihvelin, 2022; see Widerker & McKenna, 2003). The same holds for species of 'soft' or 'semi' compatibilism that argue determinism can align with moral responsibility, even if not with free will (Fischer & Ravizza, 1998, 2000; Pereboom, 2014, p. Ch 3). If we accept such objections or reject the compatibility of determinism and free will, the challenge the God Machine presents to the 'freedom to fall' becomes less convincing.

Finally, I would note what we have considered about the mind, from a scientific sense, calls into serious question the very *metaphysical plausibility* of distinguishing between choice and action—which I think we need to accept as possible for the purpose of the God Machine. The mind is characterized by emergent and irreducible properties, such as nonlinearity, and it is enacted through continual cycles of action and perception.

Assuming that interference with any form of action can isolate causal influence across the complex system of the mind and discourse of minds seems doubtful, considering the emergent nature of our mental system. Analogous to the view the 'skin-and-skull barrier is a relevant ethical watershed' and 'involves bad metaphysics' (Heersmink, 2014), it could be said that the God machine similarly draws an artificial line to separate the mind and enacted mental states, embedded rationality between choice and action, and discussions of 'metaphysical access' (Fischer, 2006b). Interestingly, adherence to such a view might, itself, trace the sorts of distorting biases that lead us to favour causal attributions in discussing morality—of the sort we explored in previous chapters (Greene, 2013; Persson & Savulescu, 2011b, 2012, 2015).

Determined Action and Moral Virtue

Second, putting aside the philosophical stalemate and related issues, even if we accept that the God Machine aligns with 'moral responsibility,' it does not fully address concerns about virtue. Virtue, much more than decisions and actions, is cultivated over time through both choice

and action (Adams, 2006, p. 3; Aristotle, 2009; MacIntyre, 1981; Snow, 2010; Swanton, 2003). As Aristotle notes: "Virtue... being of two kinds, intellectual and moral, intellectual virtue in the main owes both its birth and its growth to teaching... while moral virtue comes about as a result of habit" (Aristotle, 2009/ 1103a.14-1103b.25). This suggests that virtue requires both choice (to form the habit) and action (to cultivate the habit).

One might nonetheless reject Aristotle's suggestion that virtue requires action, turning to 'soft compatibilism,' which argues that moral responsibility can exist within deterministic constraints (Fischer, 1994, 2006a; Fischer & Ravizza, 1998, 2000). This perspective postulates that moral responsibility can be maintained despite external controls and perhaps allows some room for the cultivation of virtue even under the watchful eye of the God Machine—especially if we assume it only interferes with a limited number of seriously immoral actions. This provides an interesting perspective, but it may not fully address the complexities of virtue cultivation.

The God Machine presents a unique conundrum in this context. While effectively enforcing moral actions, it could potentially hinder the evolution of moral desires and intentions, elements that are sometimes seen as central to true moral agency (Arpaly, 2002; Vargas, 2013). Essentially, individuals might behave morally under the machine's guidance but still entertain immoral desires and intentions. This discrepancy presents a complex quandary about the nature of virtue cultivation and its compatibility with determined action.

Additionally, the constant regulation of the so-called 'God Machine' could potentially lead to moral stagnation. By enforcing uninterrupted moral correctness, the machine may inadvertently hamper the development and maturation of our moral character. This issue presents a significant challenge, especially within the context of virtue ethics.

A possible response is, again, that virtue might be achieved within a framework that provides only limited restrictions to freedom. But returning to metaphysical concerns about superficial boundaries between choice and action, I am not confident that, at this stage, we could not limit potential risks to isolated agents or a certain class of actions.

The notion of fostering virtue within certain deterministic circumstances, such as those orchestrated by the 'God Machine,' warrants further investigation and theorization—and ultimately falls in the realm of metaphysics and the broader debate on moral responsibility. Yet, in this phase of theorization, I remain skeptical that depending on external factors—like the 'God Machine' or its more realistic counterparts in neurointerventions—would not inhibit the organic

cultivation of moral virtue across various domains. This could have repercussions on the achievement of complete virtuous personhood.

Libertarian Free Will

Third, even setting aside concerns about source compatibilism and the compatibility of determined action with virtue cultivation, the God Machine faces another challenge. There are various competing theories of free will. Many accept determinism, that it is incompatible with free will, and that we still have free will. Such 'libertarian' thinkers argue that genuine free will cannot coexist within a deterministic universe. Some libertarian accounts of free will introduce promising scientific angles for understanding human freedom.

An intriguing perspective, which I believe holds promise, suggests the role of quantum indeterminism in modulating brain networks like the default mode network. These networks are associated with rational, deliberate, and intentional processes (Tse, 2013, 2018). Another school of thought investigates the cascading effect of 'self-forming actions' over time (Kane, 1998; Mele, 2006; Mele & William, 2009). These theories echo our previous discussions about human rationality, offering a fresh perspective on the dynamic interplay between quantum processes, brain networks, and human decision-making.

Among these, a particular class of 'agent-causal' libertarian theories suggests that free will and free decisions stem from an agent's ability to instigate a causal event, deviating from preceding causes and paving a path for human freedom (Alvarez, 2009; Lamb, 1977; Steward, 2009). These theories often differentiate between actions and events or things we do versus things that happen to us. To truly be free, an agent must 'author' the event (Davidson, 2001a). However, some contest that such actions might not only be 'unfree' but may not count as actions at all (Brent, 2020; Steward, 2012; Taylor, 1966; Vihvelin, 2022). They view them as occurrences rather than actions. Consequently, these theorists propose that agent causation necessitates alternatives (Clarke, 2003). This suggests, therefore, that moral responsibility and moral virtue might also require alternatives.

Assuming these theories are accurate and freedom is rooted in an agent's actions, then any external interference or disruption with these actions—such as manipulation by the 'God Machine'—could impede libertarian freedom. This, in turn, could affect moral responsibility and the nurturing of virtue. Especially with regard to virtue, the emphasis in many of these theories

is on the evolution of self-forming actions over time. This holds, I think, even if we accept responsibility and virtue were otherwise 'compatible' with deterministic restrictions.

Considering the concepts of rationality I have previously discussed, I believe it is implausible to dismiss out of hand the potential impact of interference in one domain without acknowledging possible effects on others. Further, these discussions also touch upon certain metaphysical concerns relating to the distinction between 'choice and action.' Given the recent advancements in neuroscience and the need to account for emergent properties, non-linearity, and integrated rationality, I find this distinction increasingly unsustainable.

That said, libertarian theories of free will, including agent-causal theories, face numerous objections, highlighting the ongoing and profound debate surrounding free will and moral responsibility in metaphysics (Levy, 2011; Mele, 2006; Pereboom, 2009, 2014; Strawson, 2010). While these discussions cannot be exhausted here, it is my conviction that the current version of the God Machine falls short in dispelling the intuition that neurointerventions, which disrupt our ability to decide and act, undermine the necessary freedom for moral responsibility and virtue cultivation. Striving for human flourishing manifests through voluntary decisions rather than coerced actions; we may argue that restrictions to metaphysical freedom and depriving any person of the freedom to fall jeopardizes something profoundly valuable.

This holds, I think, at the highest level of idealized theorizing, and as such, *in principle*, we cannot rule out a plausible moral asymmetry between direct and indirect interventions—between neurointerventions and conventional forms of punishment. That neurointerventions pose distinct threats to human freedom—what I take to be 'mental freedom'.

But I do not think this says enough. This is because we necessarily acknowledge we must place restrictions on freedom, particularly when we discuss crime. This leads us to a more interesting question, and what I see to be the most formidable challenge the God Machine poses to the 'freedom to fall.'

Restraining the Hidden Demons—Why Value the Freedom to Do Wrong?

Even if we find the God Machine interferes with human freedom and hinders the pursuit of virtue, it motivates another question that particularly resonates when we situate the issues in the context of criminal justice practices—"Why does the freedom to fall matter?" (Pugh, 2019, p. 75). Why should we value the freedom to commit seriously immoral actions?

We need to ground this issue in the current debate. At the highest level of ideal theorizing, the key question is whether there is a fundamental difference between these interventions in terms of their moral implications. In my argument, I have posited a *moral asymmetry* based on the potential threat that neurointerventions pose to human freedom. The notion of the freedom to fall has been invoked to highlight this risk. However, the God Machine motivates a second challenge. The question is the inherent value of a limited form of the 'freedom to fall': the freedom to commit a narrow class of seriously immoral actions—"falling big time" (Savulescu & Persson, 2012, p. 413).

I think it might be possible to frame the argument as follows. If we determine that this specific aspect of freedom lacks intrinsic worth, it may undermine the *moral asymmetry* previously established. While an in-principle difference between direct and indirect interventions may exist, it is not morally relevant, as our first-order judgments do not provide sufficient *moral reasons* to protect against interference.

I admit the God Machine, at least at first sight, poses a formidable challenge on this front. Especially when we consider criminal justice practices, and universally immoral acts, such as murder, rape, and torture, some theorists propose the use of neurointerventions that "narrowly target particular types of choice or act" to prevent individuals from performing these crimes (Ryberg, 2020, p. 63). For instance, Degrazia suggests a brain-implanted device that could inhibit sexual offences, arguing that losing the freedom to commit such acts is "no great loss" (DeGrazia, 2014, p. 63; Ryberg, 2020, p. 63).

To say this form of freedom is worthy of protection, would we not, in Dostoyevsky's terms, be suggesting that there is value to 'breaking the chains' and liberating the hidden demons of human nature (Dostoyevsky, [1880] 2003). Instead, is it not desirable to be free in thought but restricted in action, as Persson and Savulescu support, where individuals can be "as free as humans can ever be" and that "everyone is much better off for the absence of evil" (Persson & Savulescu, 2012; Savulescu & Persson, 2012, p. 413). We need to not choose either to 'reign in Hell' or 'serve in Heaven; (Milton, [1667] 2014 Book I, lines 258-263). Rather, we can relegate the darkest demons to the latter while preserving near-perfect freedom in the former. At first sight, surely this is an inebriating vision of freedom. Far more appealing than one that not only allows, or in fact, *values*, the ability to act wrongly.

I begin with this. If the question, or challenge, is whether or not we should value allowing persons to commit heinous acts, the answer is clear: of course not. Neither did Milton imply as much in his conceptualization of the freedom to fall: "Know that to be free is the same thing as to be pious, to be wise, to be temperate and just, to be frugal and abstinent, and lastly, to be magnanimous and brave" (Anderson, 2010; Milton, 1667; Milton & Hughes, 1957, p. 837; Savage, 1977, p. 286). If this is the real question (and to be charitable, I do not believe so), it is uninteresting and philosophically trivial.

Following the 'God Machine' thought experiment, it seems to me that the challenge predominantly pertains to a specific subset of freedom and a certain type of interference. The critical question is, even if this thought experiment poses challenges to moral asymmetry in a limited set of scenarios, to what extent can it generalize? How can it guide ethical discussions across different domains and, ultimately, influence criminal justice policy? This ties back into issues about the scope and permissibility of punishment.

And on this front, even within the realms of science fiction, the 'God Machine' concept—though alluring in its simplicity—proves insufficient on many fronts. It fails to shed light on pertinent philosophical issues relevant to criminal justice practices. These deficiencies range from the theoretical and conceptual to the normative and practical realms.

The God Machine as Ultimate Authority on Morality

First, at the level of ideal theorizing, the God Machine encounters significant conceptual challenges we touched on in Chapter 2. While it claims to target a specific class of universally wrong acts, such as murder, rape, and torture, it oversimplifies the complexity of human morality by assuming a clear distinction between moral and immoral actions. This overlooks the practical reality that there is often no consensus on moral principles or guidelines for determining right or wrong. The study of morality is a divisive and multifaceted field, sparking passionate debates across disciplines and social groups. And outside a narrow class of cases, as Milton emphasizes, "The mind is its own place, and in itself, can make a heaven of hell, a hell of heaven" (Milton, [1667] 2014, p. 254). This poses significant issues in the real world, tracing the 'yardstick' objection we considered earlier (DeGrazia, 2014, p. 364; Earp et al., 2018, p. 168; Ryberg, 2020, p. 57; R. Sparrow, 2014, p. 22).

Outside of this extremely narrow domain of overlapping consensus, the notion of a "God Machine" as the ultimate moral authority becomes difficult to substantiate. This difficulty is present in the realm of ideal theorizing, spanning a moral penumbra and facing devastating challenges even on the remote boundaries of broader normative structures and discourse in which neurointerventions would find any footing in the real world.

For example, recall overcriminalization, explored in Chapter 2, and the tragic case of Alan Turing we considered in Chapter 3. (Hodges, 1983; Liberto, 2018, p. 196; McTernan, 2018a; Ryan, 2020, p. 272). Consider more intermediary cases we discussed in that chapter, and evolving social *moral* and *ethic* views, and even setting this aside, the fact that what is 'criminal,' in itself, a product of political decision making—there is a distinction between what is considered moral and what is legal.

In the absence of consensus and with a proper understanding of the embedded nature of the mind, neurointerventions have the potential to profoundly impact individual psyches and the broader community of minds. This interference with unique freedoms can have far-reaching causal effects and may restrict dissenting voices. Drawing from the lessons of history, such as the case of Turing, we must consider the implications of powerful interventions that can undermine freedom, alter the collective psyche, and change society—and the very discourse in which views about morality and criminality are structured.

Morality and Complex Social Practices

Second, building on these themes, for the God Machine to function as a comprehensive moral authority, it would need not only perfect knowledge of brain-based operations but also omniscient awareness of the broader physical, social, and normative contexts. However, as Dworkin reminds us in relation to morality, "there is no algorithm for finding the truth" (Dworkin, 2011, p. 58). But even assuming such knowledge was attainable, the God Machine would require an algorithm capable of accounting for the intricate interplay of emergence, nonlinearity, and self-organization in both the mind and the environment. It would need to go beyond mere brain functions and consider the complexities of the mind and its interaction with embedded physical and social environments (Earp et al., 2018; Wudarczyk et al., 2013).

For example, depending on facts about the world, the act of killing could be seen as wrong in some instances—vengeance, retaliation, anger—but an act of necessity or compassion in

others—to alleviate suffering based on the autonomous medical decision of another. But such a determination might, for example, require accounting for constraints on 'autonomous' decision-making on the part of a patient and, in turn, a corresponding metric for autonomy, attentive to the patient's own narrative, life goals, aspirations, and so forth. What would the algorithm be for distinguishing and resolving these issues? This returns us, I think, to issues about the scope and conditions for autonomy and *freedom* and all sorts of related considerations. This is but one example among many that could be foreseen.

Moreover, particularly in the sense of virtue cultivation I have endorsed, freedom is inherently bound to the practices, narratives, and traditions of a given community (MacIntyre, 2013). Given the multifaceted aspects of human nature and morality, expecting a machine to discern good and evil actions in real-time is highly improbable, even in hypothetical terms. Based on these considerations, I believe we have reasons to question not just the conceptual and practical feasibility of targeting a vast majority of 'immoral' behaviours but also their metaphysical validity. This connects back to the belief that a significant metaphysical divide exists not just between choice and action but also in the recursive relationship involving the social discourse in which both are embedded.

Practical Considerations, Bias, and Final Thoughts

Another issue with the God Machine is one that applies to Frankfurt-style experiments in general. Specifically, it may be too abstract and removed from everyday experiences to provide practical solutions to real-world moral problems (Speak, 2002; Vihvelin, 2022). This poses issues when we consider how far it might generalize into practical domains for discussions about punishment.

The critical question is this: does the concept of an 'all-knowing God machine' help unravel the intricate links between complex matters such as crime, the labyrinth of mind-morality connections, and the promise and perils of neurointerventions? We must remember our past definitions of 'criminal' and 'immoral' have sometimes been erroneous. Therefore, branding individuals as 'criminals' from the outset, as was done with Turing not long ago, may not fully represent these complex themes. This thought calls for profound introspection.

Building upon the themes discussed thus far and the concerns raised about ideal theorizing, it would be a profound understatement to merely suggest that the God Machine

overlooks a multitude of crucial practical issues with significant moral implications (McTernan, 2018a, p. 284; Vallentyne, 2018b, p. 138). Respectfully, and so as not to cast aspersion, I would simply note concerns about how thought experiments are "often abused, though seldom deliberately" and risk being used as pedagogical and rhetorical devices which are too removed from actual possibility "to offer any illuminating perspectives for ongoing research" (Behme, 2013, p. 377; Bokulich, 2001, p. 304; Dennett, 1984, p. 24). There is nothing inherently objectionable about the 'God Machine.' The risk of misuse, I think, would be if it is deployed to offer any practical guidance in issues in the real world.

In short, I believe the "God Machine" falls short on practical, conceptual, and theoretical grounds. It oversimplifies the complex realities of human behaviour, undermines the fundamental freedom of choice, disregards the complexities of moral discourse, and is based on problematic metaphysical views of human rationality and decision-making. At the risk of repetition, I would conclude this exercise by returning to discussions about ideal theorizing inherent throughout this dissertation.

From Ideal to Practical Theorizing

So, as we conclude our foray into the highest realm of ideal theorizing, I think some of the larger issues come into clearer focus. Recall Nietzsche's declaration, "God is dead. God remains dead. And we have killed him… Must we ourselves not become gods simply to appear worthy of it" (Nietzsche, [1882] [1974])? Perhaps we might see the God Machine as some revived deity to fill a divine void with technological intervention. It is a strange and powerful notion, one that echoes within the vaults of our technological aspirations and ethical conundrums. Yet, the reality is any such embodiment will be a reflection of our own image and a product of our design. It seems foolhardy to think that the same 'evils' we wish to rectify in others could not potentially be reflected in such a creation—or at least, I find it difficult to truly believe this is the case as matters now stand in criminal justice practices.

But the 'freedom to fall' is not immune to similar criticisms. Biblical notions of unfettered freedom and lofty aspirations of flourishing do not necessarily trace onto complex issues about crime nor the sorts of capacities required to realize human virtue. As Kierkegaard suggests, "Life can only be understood backwards; but it must be lived forwards," but "To dare is to lose one's footing momentarily. Not to dare is to lose oneself" (Kierkegaard, 1849/1980).

Just as in the story of Adam and Eve in the biblical Garden of Eden, the freedom to fall has been an intrinsic part of our understanding of what it means to be human. And yet, the reality of being human—and all too human—is far more nuanced, as our choices and moral actions are often constrained and influenced by many complex factors, and in a community, neither unqualified nor unbound.

But in closing, I maintain even at the highest level of analysis—esoteric thought machines, metaphysical notions of freedom, morality, and virtue—there remains a *moral asymmetry* between direct interventions and conventional forms of punishment. This moral asymmetry is based on the fact that neurointerventions, *in principle*, pose distinct threats to metaphysical freedom, key facets of our moral life, and this generates moral reasons to extend protection against such interventions that do not extend to conventional forms of punishment.

This exercise serves to identify valuable lines of inquiry that will guide discussions that follow. However, the task at hand, and the most important one, I think, is to explore how this concept of 'freedom' permeates outside extreme cases, within the realms of political philosophy, moral theories, and ultimately, applied disciplines like public policy, biomedical ethics, and law. This exploration will continue in the remaining parts of this chapter.

Freedom, Morality, and Domains of Inquiry

We have considered freedom, in the most fundamental sense, as a *metaphysical* concept, with reference to diametrically opposed thought experiments. But this is far from exhausting the moral or normative landscape. As such, I will reground discussion as we move away from the ideal normative and moral conversation towards non-ideal, practical normativity.

Morality and Normativity—The Landscape

The remaining portion of this chapter aims to explore notions of moral and normative *asymmetry* that stand in opposition to assertions of punishment *equivalence*. But we are considering these in a specific context, to which we must be attentive at all junctures.

By their inherent nature, traditional forms of punishment involve burdens imposed by the state and treatments we might conventionally perceive as 'wrong.' For instance, incarceration deprives individuals of their liberty, severs relationships, and impedes the pursuit of life objectives.

Similarly, mandatory brain interventions—forcible medical interventions deployed to alter morality—intuitively seem wrong. I introduced the 'freedom to fall' and 'God Machine' to illustrate that the sort of 'wrong' neurointerventions imposed are of a different kind because they pose distinct threats to human freedom at the *metaphysical* level and an extreme end of ideal theorizing. This addresses profound questions about what sort of freedom is *required* to be moral—or even make sense of the term.

But even accepting such an understanding, further exploration leads us into complex realms of 'morality' and associated debates on 'normativity' (Brink, 2010; Dancy, 1993, 2023; Kamm, 2007; Mackie, 1977; Wedgwood, 2007, p. Ch 1). In order to comprehend the 'wrongness' of such intrusions, establish grounds for decisions about right and wrong actions, and determine if there's a distinction between them, we need to delve into these domains. However, this exploration is fraught with challenges.

Navigating the abstract realms of 'morality' and 'normativity' opens up a Pandora's box of conceptual and even existential questions. Can morality, for instance, be deemed a universal principle or simply a human creation (Mackie, 1977)? It seems improbable to affirm normative claims solely based on brain features or evolved social behaviours (J. D. Greene et al., 2001, p. 2107; Moseley, 2020; Wheatley & Decety, 2015; Wiseman, 2016). In other words, it's challenging to map neural facts onto judgements of moral right or wrong.

Further, beyond the realm of abstract theorizing, the intricate tapestry of normative landscape resists simplification into a singular theory. This landscape is a web of complex moral and normative interactions across diverse domains. Disputes over moral and political values infiltrate our culture, requiring an interplay of varied theories for everyday decision-making (Dworkin, 1996; MacIntyre, 2007). And in a sense, much like the mind, what emerges is a complex emergent system which cannot be fully reduced to facts about the world or any particular theory or conceptual framework—much like the very biological substrates from which they arise.

But despite their elusive nature, morality and normativity hold pivotal roles across various areas of thought and inquiry relevant to discussions in neuroethics. First-order normative theorizing, particularly deontological and virtue-based theories, provides fertile grounds for inquiry for equivalence analysis (Gert, 1998; Singer, 2011). Equally, political philosophies, such as Rawls's theory of justice, apply ethical principles to tackle societal challenges (Freeman, 2007; Rawls, 2001, 2005, 2009). Law, public policy, and applied ethics are related fields guiding

the interpretation of legislation and policy decision-making (Dworkin, 1985, 1986; Thacher, 2006).

Just as identifying a core in criminal behaviours aids our understanding, it is not unreasonable to approach moral and normative theories similarly. And perhaps it would be sufficient simply to conclude that we should merely accept or take for granted that *freedom* is valuable. But I do not think this goes far enough. More is required when we address difficult issues about the brain, mind, punishment, and forcible intervention.

Certain elements, such as human freedom and rationality, repeatedly surface across these theories. But various conceptions offer different insights and provide valuable areas for inquiry. We cannot explore them all here. The goal is not to identify every challenge nor entertain every response. It is to highlight that concerns about 'asymmetry' grounded in human freedom generalize in powerful ways across various domains, further posing challenges for equivalency arguments—and perhaps offering suggestions for future research.

Equivalence, Asymmetry, and Freedom as a Unifying Theme

Central themes of human freedom and rationality have implications across diverse areas such as ethics, political theory, and law. But outside this domain, conversations of freedom are fragmented, reflecting various iterations with an accompanying flourish.

The intoxicating vision of human freedom "unfettered by the given" (Sandel, 2007, p. 99). The power to shape our lives, 'in pursuit of the things that we value' and the aspiration to be the "undivided author of [one's] own life" (Habermas, 2003, pp. 89-90; Stemplowska, 2018, p. 343; Tadros, 2011, p. 130). Perspectives at the core of Kantian morality, placing paramount importance on a 'self-determined life' seen as an end in itself, emphasizing that freedom is a prerequisite for morality—"morality presupposes freedom" (J. D. Greene et al., 2001, p. 2107; Moseley, 2020; Wheatley & Decety, 2015; Wiseman, 2016).

The same intuitions find footing in diverse placeholders such as 'rational agency,' 'moral agency,' 'liberty,' 'autonomy,' and 'self-determination' recur in ethical discourses. However, these complex concepts can be oversimplified or misunderstood when applied universally. Hence, it's crucial to understand 'freedom' in relation to its specific context.

Therefore, for the balance of this chapter, I investigate how the foundational concept of 'freedom' is interpreted across different fields and its implications therein. We will analyze this

spectrum, from ideal to practical theorizing, identifying potential asymmetries in morality and normativity. We will focus on three areas: basic moral theorizing, political theory and philosophy, and applied fields such as public policy, applied ethics, and law.[108] At each domain, I identify how leading objections to neurointerventions might bear an important connection to human freedom—and what this suggests about equivalence and asymmetry.

First-order moral Theories

Starting with first-order moral theorizing, a range of theories can guide our judgements about actions' morality and bolster arguments about the moral asymmetry between direct and indirect interventions—casting light on the importance of rationality and 'human freedom.'

Our exploration primarily focuses on deontological theories, which have received significant attention in the neurointervention discourse (Heinrichs & Stake, 2018, p. 163). These theories, emphasizing "rightness," are steered by universal laws (Kant, [1785] 2002). Another pivotal approach considered above grounded 'virtue theories,' highlighting the significance of fostering virtuous character traits and moral virtues, and what "human beings ought to be and what sort of life we ought to lead" (Anscombe, 1958; Foot, 2002; Rüther & Heinrichs, 2019, p. 356). Moreover, consequentialist considerations (Singer, 1972), which might be seen as generally incorporated into penal theories justifying punishment, are inherent in necessary restrictions to freedoms in the realization of larger collective goods and are discussed in this context throughout what follows.

Demarcating first-order theories is useful as it helps in understanding the foundational ideas that drive various moral and ethical theories. This initial level of theorizing often provides the basis for more complex or practical applications in various fields. This proves instrumental in discerning moral asymmetry when considering compulsory neurotechnologies that could potentially impinge on rationality and human freedom.

First-Order Conceptions of 'Freedom'

In the realm of first-order theorizing, the notion of 'freedom' and its inherent value is justified in different ways and sometimes simply taken for granted.

Central to first-order moral theorizing and particularly, Kantian ethics, is the veneration of a "self-determined life", perceived as an "end in itself" (Kant, 1999 (1781); Lee, 2012, p.

338; Stolzenberg, 2008; Varga, 2015, p. 74). This philosophy embraces the premise that "morality presupposes freedom" (Ivison, 2007, p. 97). Such thought aligns with the understanding of rationality as an inherent or unique human trait. It echoes the idea of rationality as a 'natural' or 'distinctive' characteristic of human beings (Hauskeller, 2011, p. 76; Sandel, 2007, p. 45), perhaps with root in notions of 'self-authorship' and dignity reflected in our earliest pre-hominid ancestors (Foster, 2015). In turn, this culminates in distinctive rational capacities necessary for participation in a moral community, reinforcing norms of respect within societal practices (Shaw, 2014, 2018; Varga, 2011; (Fukuyama, 2003, 2011). As previously argued, theories of virtue and the fostering of virtue depend on habitual action and, therefore, 'freedom'— the liberty to make choices and act—in the quest for practical wisdom and human flourishing (Adams, 2006, p. 3; Aristotle, 2009 1103a.14-1103b.25; MacIntyre, 1981; Snow, 2010; Swanton, 2003).

The question is how neurointerventions might represent threats to these general notions of 'freedom' and whether they are equivalent or different, *in principle*, to those posed by conventional forms of punishment. There are a few lines of argument in the ethical discourse that cast light on this issue.

Arguments from Human Nature—Naturalness, Hubris and Playing God

First, within first-order normative theorizing, to identify moral asymmetry, one might turn to 'arguments from human nature,' which are prominent in both deontological theory and virtue ethics. Critics assert that neurointerventions represent a form of 'playing god,' potentially eroding essential aspects of human nature and our unique relationship with the universe, fueling fears of artificially constructed moral agents and a 'posthuman future' (Erler, 2020, p. 384; Fukuyama, 2003; Glendinning, 1990, p. 84; Habermas, 2003; Kass, 2003; Rüther & Heinrichs, 2019, p. 174; Sale, 1995; Sandel, 2002; Sandel, 2007). An unbridled drive to master and control human nature might be viewed as hubris, infringing on the virtue of humility and an appreciation for our innate abilities and distinctive way of being, in favour of a 'Promethean project' of mastering nature and instead of embracing life as a gift (Hauskeller, 2011, p. 76; Sandel, 2007, p. 45).

But the definition of 'naturalness' and 'human form' can be problematic, theoretically, and politically divisive, and risks being motivated more by visceral reactions and 'status quo

bias' than rational thought (Bostrom & Ord, 2006; Bublitz, 2020b; Buchanan, 2009; Buchanan, 2011; Caviola et al., 2014; Dresler et al., 2019, p. 1142; Green, 2010; Moore, 1903; Persson & Savulescu, 2012, p. 115; Rüther & Heinrichs, 2019, p. 170). Even outside of these concerns, in the realm of ideal theorizing, the problem is that notions of 'the natural human form' can also be relative, given human diversity and varying conceptions across cultures, societies, and even within the same culture or society, over time (Choudhury & Slaby, 2016; Conrad & Leiter, 1981; Fox & Swazey, 2013).[109] Meanwhile, extending protection to the 'natural form' and finding the "wrong generating" features of neurointerventions can be challenging. We need to justify why altering 'human nature' would be morally wrong (Buchanan, 2009; Buchanan, 2011; Groll & Lott, 2015; Kahane et al., 2017; Lewens, 2012).

Notwithstanding, carefully refined metrics based on 'human form' show at least *some* promise. I think it might be open to adopting a "minimal species ethic as a formal notion of the good" (Varga, 2015, p. 81) or perhaps to ground unique features that permit our participation in a moral community (Habermas, 2003; Fukuyama, 2003), grounding normativity in the form of rationalist procedural morality ethics proposed by theorists like Habermas and Rawls (Habermas, 1984, 1990, 1996a, 1996b, 2003; Rawls, 1955, 2005, 2009; Varga, 2015, p. 81).

Even so, such a conception does not inherently create asymmetry between neurointerventions and conventional punishment. Both criminal behaviour and punitive responses might be seen as 'unnatural' in their own rights. Furthermore, our irrational justice system and the limits of our rationality, which are hampered by widespread indirect subversion, also call into question whether 'rationality'—as we commonly conceive it—could be seen as 'natural' to the human form.

However, while I accept arguments from 'naturalness' and 'human nature' may be problematic, I argue they are worth exploring to understand the potential moral objectionability of neurointerventions. If we accept a minimal description of the human form comprising vital aspects of rationality that allow participation in the moral community, building on themes in the last chapter, this could include distinctly human emergent mental capacities, enabled by introspection and self-organization, that extend into wider social and normative landscapes, allowing us to engage in a larger social discourse, possibly generating human dignity and moral value. On this basis, neurointerventions that disrupt these rational capacities in distinct ways could be morally objectionable.

Moreover, while arguments about 'human nature' and the 'human form' may initially appear dubious, I think they might be strongly bolstered if grounded in related notions of 'virtue cultivation,' which we have discussed, where such capacities might be seen as a prerequisite for attaining practical wisdom in pursuit of flourishing (Adams, 2006, p. 3; Aristotle, 2009 1103a.14-1103b.25; MacIntyre, 1981; Snow, 2010; Swanton, 2003).

Finally, in response to claims that a growing understanding of the limits of our rationality cast doubt on features which are truly 'natural' to the human form, I would reiterate my claim in the last chapter of the significance of our ability to recognize these limits—perhaps the most distinctive human feature—and in so doing, we find ourselves echoing Darwin's rhetoric: "our love of nature and life—our love for what we formerly conceived as "God's creation—can outlive our dying belief in God" (Darwin, 1860/2010; Davies, 2009, p. 4).

While acknowledging that far more work would be required, what is crucial for our purposes is that 'human rationality' and 'human freedom' are important in assessing whether appeals to the 'human form' raise asymmetry.

Human Dignity, Respect-Based, and Communicative Theories

Further, in deontological critiques of neurointerventions, respect-based and communicative theories feature prominently—and all share some important connection to human rationality and human freedom. These perspectives pivot on the moral principle that actions showing respect are virtuous, whereas those indicating disrespect are inappropriate. This principle has strong roots in Kantian philosophy, particularly the imperative to "treat humanity in every case as an end withal, never as means only" (Kant, [1785] 2002).

Respect, recognized as "a weighty value—a kind of trump card—amongst moral values", is closely associated with human dignity (Beauchamp, 2013; Bennett, 2018, p. 265; Dworkin, 1988b; Focquaert et al., 2020, p. 140; Foster, 2011; Gilabert, 2019; Holmen, 2020; Lewis, 2021, p. 17; Matravers, 2018, p. 81; Tonry, 2016). However, respect is not innate; it arises from mutual relationships recognizing others as deserving of moral respect (Bennett, 2018, p. 260; Habermas, 2003, p. 33; Varga, 2015, p. 81). Moreover, the social meaning of an action, including whether it exhibits respect, is contextual and culture-dependent (Hampton, 1991, p. 1670; Shaw, 2018, pp. 322-323). Further, as we will see, the importance of respect and dignity

widely ripples into fields like biomedical ethics and legal thought (Beauchamp, 2013; Dworkin, 1988a; Focquaert et al., 2020, p. 140; Holmen, 2020; Lewis, 2021, p. 17)

Respect-based theories also have a close relationship with human agency and autonomy, valuing individuals as rational beings (Varga, 2011, p. 75). As Lippke argues, this ties to the imperative that demands us to "respect the status of persons as rational beings who are capable, within limits, of making choices about the courses their lives will take." (Lippke, 1998). Stephen Darwall calls this 'recognition respect': "[people] are entitled to have other persons take seriously and weigh appropriately the fact that they are persons in deliberating about what to do" (Darwall, 1997, 2006; Darwall, 1977, p. 38). Lippke (1998) argues satisfying this entitlement requires that we "respect the status of persons as rational beings who are capable, within limits, of making choices about the courses their lives will take."

Critiques arise when biomedical enhancements, like neurointerventions, are seen to disrespect agency and generate asymmetry between the "designed" individual and their "programmer" (Habermas, 2003, pp. 64-65). Compulsory neurointerventions might signify to offenders that they are less deserving of respect, undermining the 'communicative ends of punishment' and there is a risk they communicate to an offender they are not moral equals but a "broken machine to be fixed," a 'puppet' or 'something less than human' (Bomann-Larsen, 2013; Olivia Choy et al., 2018, pp. 34-35; Duff & Duff, 2001; Duff & Hoskins, 2019; Shaw, 2012; Sparrow, 2013, p. 91). This aligns with C.S. Lewis's critique of shifting from retributive to paternalistic justice models. Lewis cautions that viewing offenders as merely products of external factors, requiring paternalistic 'repair,' risks transforming the justice system into one that disregards individuals as moral agents, thereby eroding the foundational principles of moral blame and responsibility.(Lewis, 1953)[110]

However, respect-based accounts are subject to various challenges. For example, some identify the concept of 'moral respect' as somewhat ambiguous (Pugh, 2018, p. 111). Others argue that neurointerventions could be justified in principle as they convey disrespect proportionate to the offender's violation, similar to incarceration (Douglas, 2014b; Bullock, 2018, pp. 159-160). Moreover, if neurointerventions can theoretically enhance an offender's rational capacities, some suggest they might be seen as actually *fostering* respect rather than diminishing it (Holmen, 2020). Along these lines, Emma Bullock argues that neurointerventions

could sometimes be justified under the concept of 'moral paternalism' (Bullock, 2018, pp. 159-160).

But I think issues of respect become paramount in light of specific concerns about current incarceration conditions. The severe realities of prison life, historical mistreatment, and potential safety and efficacy issues surrounding neurointerventions all underscore the importance of cultural and contextual considerations. In these instances, the matter of respect becomes particularly weighty. If the very basis of punishment is to condemn certain conduct that might be seen to be morally objectionable and premised on the view of persons as rational persons, there seems to be something deeply problematic about this state of affairs and further projects that might relay a distinct, or cumulative form of disrespect.

Theories based on respect provide a crucial framework for this discussion, especially when recognizing the unique threat neurointerventions pose to an individual's moral agency. This might imply that the parity principle may not adequately account for aspects of our mental lives that remain unaffected by indirect interventions. It seems plausible, then, that concerns about human dignity and respect are closely tied to themes of rationality, moral agency, and, ultimately, the broader notion of human freedom.

I propose first-order theories provide crucial insights into the core ideas driving moral and ethical theories. They serve as the foundation for practical applications in various fields. Recognizing the unique threat neurointerventions pose to human rationality; we must thoroughly explore the landscape of potential moral asymmetry. This exploration helps us understand the impact of compulsory neurotechnologies on rationality and human freedom while ensuring the preservation of essential values like virtue, dignity, and respect.

Political Legitimacy and the Scope of State Authority

Venturing into the territory of political philosophy, we position ourselves further down the spectrum between ideal and non-ideal theory, grounding our discussion more in real-world dynamics between citizens and the state—straddling the line between moral and normative asymmetry. Building upon our previous discussions, we discover parallels to the considerations of dignity and respect for moral agency. But social and relational aspects take a more central role in connecting moral or normative aspects to the underlying political dynamics between the state and its citizens. Within this discourse, the concept of 'political legitimacy' becomes highly

relevant in analyzing freedom and the implications of neurointerventions. This is particularly the case in the context of punishment, which involves state-imposed burdens and the structure of relationships between citizens and the state.

Political Conceptions of Freedom

In the realm of political philosophy, freedom and rationality underscore longstanding notions of the "equal Right that every man hath, in his Natural Freedom (Locke, 1824b II.54), rights to personal autonomy and 'freedom of thought' (Berlin, 1969), treating citizens as free and equal (Rawls, 2005, p. 337), as 'authors' (Habermas, 2018), reflecting the consent of the governed (Hobbes, [1651] 1980; Locke, [1651] 2016) moral capacities (Rawls, 2005, p. 337); the very legitimacy of the state is a corollary of autonomy (Locke, 1824b, [1698] 1998; Nozick, 1974). In Hobbes's view, this freedom reflects the view humans are both the 'artificers' and 'matter' of the institutions that oversee social and political arrangements (Hobbes, [1651] 1980; Ivison, 2007, pp. 13-14). Society, then, becomes "a community of minds negotiating and arguing over ideas, rules, [and] values" (Powell & Derbyshire, 2018, p. 358). Terminology in this domain often focuses on general notions of 'political freedom,' 'liberty,' and 'autonomy' (Berlin, 1969). In this context, 'political' freedom might be understood in various ways, and so too, the threats neurointerventions might pose.

Political Legitimacy

Political legitimacy, in essence, concerns the rightfulness of a state's power and the equilibrium it maintains with individual liberties. On many plausible accounts, the foundation of the state and its legitimacy lies in the form of a social contract, where citizens surrender certain freedoms to a central authority in exchange for protection and order (Hobbes, [1651] 1980). Its ongoing existence hinges on the 'general will' or collective consensus of the citizenry (Rousseau, [1762] 1964).

However, political legitimacy has been interpreted differently; for example, the state's respect for citizens' rights to life, liberty, and property (Locke, [1651] 2016), the 'moral defensibility' of state actions (Beetham, 2013), fairness and equitable distribution of primary goods (Rawls, 1971, 2005), and more generally, a shared belief in the government's right to rule

and consent to its power. The law's addressees can also view themselves as its authors, as part of a deliberative democracy and communicative action (Habermas, 2018).

As I see it, this general inquiry for political legitimacy involves identifying the moral and normative properties ascribed to the structured dynamic relationships between sovereign and subject. And the notion is very closely tied to themes of freedom I have identified, in addition to those traced through first-order theorizing—mutual recognition, human dignity, and respect for rational agency, all underpinning a rationalist procedural morality, most closely reflecting those proposed by theorists like Habermas and Rawls (Habermas, 1984, 1990, 1996a, 1996b, 2003; Rawls, 1955, 2005, 2009; Varga, 2015, p. 81).

Neurointerventions, Punishment, and Political Legitimacy

Several theorists have touched upon the implications of neurointerventions for political legitimacy, but this area remains largely underdeveloped, especially concerning punishment equivalence. Given concerns about parity and justifications for punishment, it becomes crucial to address the question: should we be concerned about the compulsory administration of neurointerventions for criminal offenders? This is a challenging inquiry.

The state bears the responsibility to address crime, prevent harm, and ensure the well-being of its citizens. This aligns with various conceptions of legitimacy and theories of punishment. Punishment serves to deter crime, protect society, and rehabilitate offenders, reflecting the principles of the social contract, such as 'protection and order' (Hobbes, [1651] 1980), and the state's mandate for safety and equitable distribution of societal goods (Rawls, 1971, 2005). Justifications for state authority, including the 'harm principle' and 'moral paternalism,' may limit individual liberties and override autonomy in specific cases to prevent harm and promote offender well-being (Dworkin, 2005; Mill & Mill, [1859] 1966).

In this context, if safe and effective, neurointerventions that aim to reduce crime, enhance societal protection, and facilitate offender rehabilitation can align with the foundational principles of the social contract, reinforcing the legitimacy of the state. Compared to prolonged imprisonment, neurointerventions may be less intrusive, adhering to the least infringement principle in neuroethics, particularly when considering a 'moralized baseline' like minimal incarceration (Lavazza, 2018). For individuals unresponsive to traditional rehabilitation, neurointerventions could provide an appealing solution (Douglas, 2014c). If we accept that

offenders bear some liability for at least minimal incarceration, it may seem that safe and effective neurointerventions do not inherently delegitimize state administration.

Notwithstanding, discussions on the legitimate scope of state authority are complex, given the myriad of concerns we have identified up to this point.

Contingent, Practical, and Political

Let us begin by accounting for certain practical considerations. First, justifying political legitimacy through punishment presents similar issues to those previously identified with the 'Punishment Claim.' Prevailing punishment practices in Western countries, like the United States, often lack rational justification. Whether through conventional or 'neuro' and 'novel' forms of punishment, threats to political legitimacy can arise or be exasperated alongside a growing arsenal of neurointerventions: "If a state tolerates (or worse, encourages) serious social injustices, then this may undermine the state's standing to punish offenders who are also victims of such injustices" (Duff, 2007; Holroyd, 2010; Husak, 1992). I think discussions in Chapter 1 are sufficient to identify these issues, and I need not say more.

However, these concerns become more profound within the political domain, where considerations of legitimacy, in my view, incorporate both normative aspects and non-ideal real-world facts. In this context, the idea of legitimacy is influenced by historical precedents of moral and medical failures in the Western world. These precedents cast a shadow of doubt over the prospective use of neurointerventions and underscore the presumptuousness of claiming we have truly assimilated lessons from our past errors. This is particularly relevant when scrutinizing our current punitive practices, a critique often directed towards societies that neglect historical lessons – us who 'forget the past' (McTernan, 2018a).

Further, in any political arrangement, manipulation of the identity of persons could lead to public concern and political instability, more so if interventions are manipulative or risky and thus could undermine state trust (Foucault, [1975] 2012; Hardin, 2002), particularly given questions of who would serve as the ultimate moral authority given the divisive nature of morality (Bennett, 2018, p. 260; Shaw, 2018, p. 322), the need to account for cultural diversity (McCoy et al., 2020, pp. 208-209), and accommodate broader calls for recognition of neurodiversity (Garland-Thomson, 2012; Goering, 2018, p. 38).

I think there is also a further challenge. To accept punishment is justified within the structure of human societies requires situating it in a larger realm of human norms and social behaviours, tracing fundamental features of human behaviours with evolutionary roots (Banja, 2018, p. 288; Decety & Howard, 2013, p. 49; Pascual et al., 2013; Wheatley & Decety, 2015; Young & Dungan, 2012). In modern societies, punishment is an expression of one of many institutions that govern relations between individuals. But another is the notion of *rights*—being moral, political, or legal. This is an issue we briefly return to in the concluding chapter.

Explaining punishment equivalence at this level requires accounting for its role in social practices, its relationship with rights, and the broader system of distribution and allocation. This conceptual issue calls for a deeper exploration and analysis to address the complexities involved. While some scholars have begun to draw connections (Douglas, 2014c, pp. 109-110; Matravers, 2018, p. 88; Vallentyne, 2018b), the issue of determining rights and their scope further adds to the intricacies that require extensive ethical theorizing. It seems to me the penal justifications for state punishment may not necessarily align with countervailing normative reasons that can restrict the scope of permissible punishment.

Ideal Theorizing, Rational Agency, and Manipulation of the Source of Legitimacy

Let us set aside these other considerations and focus on the central concern: Do neurointerventions pose a distinct threat to human rationality and 'human freedom' and political legitimacy that traditional interventions do not? I have identified the potential threats neurointerventions pose to metaphysical freedom, which is crucial for moral ascriptions and the cultivation of virtue at both individual and social levels. Additionally, we explored the importance of respect and dignity in mutual relations as well as the moral asymmetry caused by neurointerventions' ability to circumvent rationality. I think in the realm of political theory, the state's potential to intervene in this manner, altering the moral landscape of its constituents, creates a further asymmetry of a related but distinct nature.

I argue that the sphere of liberty I have identified is closely connected to the larger political discourse, shaping meaningful moral dialogue and guiding state-citizen interactions. Revisiting our prior discussions on 'rationality,' I underscore the concept of an 'embedded' mind—an agency whose rational capacities are scaffolded into the normative environment (Clark, 2008a; Fuchs, 2004, 2008, 2011; Fuchs & Schlimme, 2009; Glannon, 2009; Pouw et

al., 2014, p. 53).[111] As social and relational organs, our brains engage in a recursive dialogue with this environment, encompassing normative properties, including political discourse (Hobbes, [1651] 1980; Ivison, 2007, pp. 13-14).

The concept of citizen 'authorship' arises from these freedoms, reflecting the consent of the governed and the citizens' moral power. In theoretical discussions, the unique ability of neurointerventions to circumvent these norms of respect and recognition in state-citizen relations is cited as a cause for concern. It undermines the membership of a moral community and the moral authority of the state (Bublitz, 2013, 2014, 2015a, 2015b, 2018; Bublitz & Merkel, 2014; Craig, 2016; Sententia, 2004; Sententia, 2013). Direct moral manipulation through neurointerventions is seen by some as exceeding the state's appropriate sphere of influence, 'manipulating the source of its own authority' and threatening the state's moral authority to punish (Bennett, 2018, p. 260; Bublitz, 2018, pp. 316-317; Shaw, 2018, p. 322).

It is important to note that the mere fact that neurointerventions 'circumvent rational capacities' does not automatically establish *moral asymmetry* at large. While this can potentially undermine political legitimacy, there are other actions of the state that also bypass rational capacities, as discussed in the neuroethical literature. As we have previously explored, the state's actions can unintentionally undermine the rational capacities of its citizens through various means, such as harmful prison environments, indirect subversion, and societal pressures referred to as dark nudges. However, the analogy with the baseline punishment equivalence remains relevant. Just because state subversion undermines legitimacy in certain areas does not mean it is justified in others.

I cannot address all of these issues. The point is that identifying the potential sources of moral and normative asymmetry between neurointerventions and conventional punishment bears importance in the realm of political theorizing. Without further theorizing, we cannot rule out the risk that neurointerventions do not pose distinct threats to political legitimacy or the risk they might reshape individuals and the societal fabric of fundamental norms and perspectives in distinct ways—in large part, based on threats to 'mental freedom' of subjects. This raises questions about their implications for political legitimacy and the complex interplay within the community of minds.

Applied Domains: Bioethics, Public Policy, and Law

Descending from the abstract sphere of first-order theorizing and political philosophy, we enter the more tangible landscapes of biomedical ethics, public policy, and law. Because the focus of punishment equivalency arguments is generally 'ideal theory' due to restrictions like the 'in principle constraint,' I only touch on this area briefly.

In this domain, the focus is not limited to *moral asymmetry*, first-order concerns, or conceptual frameworks. Instead, it is relevant to account for a potential *normative asymmetry* that is informed by moral considerations but extends further and accounts for non-ideal considerations, social norms, legal norms, and practical and prudential considerations. This encompasses a range of assumptions and considerations that inform our judgments and shed light on practical applicability and guidance in real-world situations. This opens the door to consider a wider range of considerations and potential grounds for asymmetry, responsive to the growing recognition of the need for a clearer delineation of ideal and non-ideal considerations and identification of underlying assumptions (Nadelhoffer et al., 2020, p. 196; Ryberg, 2020, p. 188). I reiterate, based on the analysis in the first two chapters, that as we enter the non-ideal realm, all of the concerns I have raised in those chapters loom large.

Freedom in Practice—Bioethics, Public Policy, and Law

In the domains of bioethics, public policy, and the law, the concept of 'freedom' holds tremendous importance and intrinsic value. It is so deeply ingrained that we often take it for granted, expressed through various terms like 'liberty,' 'agency,' and 'autonomy,' and enshrined in legal protections such as 'freedom of expression,' 'freedom of religion,' and 'freedom of thought.' Scholars even *expressly* emphasize the significance of understanding these concepts, focusing on observable capacities, democratic authority, autonomy, and rights, attempting to distance themselves from engagement in extensive metaphysical or religious debates (Dubljević, 2013, 2016). This broader perspective, which Bublitz frames in terms of a 'normative dualism,' encompasses a wide range of issues, including legal conceptions of 'freedom' that bring their own distinct considerations into the mix (Bublitz, 2020b; Bublitz & Merkel, 2014). This includes existing norms and legal protections (Bublitz, 2013, 2014, 2015a, 2015b, 2018, 2020a, 2020b; Bublitz & Merkel, 2014).

For example, bioethics is guided by principles such as beneficence, non-maleficence, justice, and respect for autonomy—the latter grounded on the significant value of 'individual self-determination' emphasizing the importance of acquiring *valid consent*[112] for any medical procedure (Beauchamp, 2013; Beauchamp & Childress, 2013; Faden & Beauchamp, 1986; Friedrich et al., 2018; Gallagher, 2018; Gillon, 2003; Glannon, 2019a; Goering et al., 2017; Pugh, 2020; Sjöstrand & Juth, 2014; Smith, 1997; Walker & Mackenzie, 2020; Witt, 2017; Zawadzki & Adamczyk, 2021).

These principles often inform the realm of public policy, where ethical insights emphasizing autonomy and individual freedom intertwine with institutional actions and carry real-world implications that have the power to alter the lives of individuals and society (Farah et al., 2014; Feigenson, 2007; Hsu et al., 2019; Joseph, 2008; Morse, 2005; Morse et al., 2013; Wolpe, 2009). One example is that, the application of neurointerventions in the criminal justice system necessitates the involvement of medical practitioners, raising critical concerns. This intersects with ongoing debates about the role of medicine in criminal rehabilitation, such as forensic psychiatry and the ethical dilemma of dual loyalty—whether physicians should participate in capital punishment or restoration of capacity to execute. These scenarios illustrate a broader risk: the potential confusion between therapeutic aims and social control, complicating the use of biomedical interventions within the framework of criminal justice (Bublitz, 2018, p. 296; Forsberg, 2018, pp. 64-65).

The law then serves as the final frontier where ideal theoretical discourse, public policy, and normative considerations intersect (Chandler, 2018; Lavazza, 2017; Nicole A Vincent et al., 2020a). Criminal and penal law function as regulatory mechanisms, maintaining social order and enforcing sanctions against harmful behaviours (Duff & Duff, 2001; von Hirsch & Ashworth, 2005). At the same time, the contemporary legal landscape places significant emphasis on human rights, which highlight the importance of concepts like human dignity and autonomy (Beitz, 2001, 2011; Buchanan, 2007b; Edmundson, 2012 part II; Morsink, 2009; Neier, 2013; Pogge, 2005; Tuck, 1979). Concepts of 'liberty' and 'freedom' are deeply woven into the fabric of the law, carrying substantial normative importance (Bublitz, 2015a; Bublitz & Merkel, 2014, p. 60; Erler, 2020, pp. 397-399; Ligthart et al., 2019, p. 120; Sententia, 2013; Wolpe, 2018).

For example, in Canada, the *Charter of Rights and Freedoms*[113] underscores notions of mental freedom, reflected in the "forgotten freedoms" of thought, belief, and opinion opinion (Laidlaw, 2024; Newman, 2019, 2021), a categorization that aligns with international human rights standards like those outlined in the Universal Declaration of Human Rights[114] and the International Covenant on Civil and Political Rights.[115] These freedoms, protected under Section 2(b) of the *Charter* alongside expressive rights, form a critical legal framework especially pertinent when considering the ethical deployment of neurointerventions within the criminal justice system. Such interventions necessitate a nuanced examination to ensure they respect individuals' autonomy and do not infringe on these fundamental human rights, offering a complex interplay with legal norms that also reflects considerations found in American constitutional protections. This legal infrastructure demands careful scrutiny to ensure that neurointerventions, while potentially beneficial, remain non-coercive and respectful of an individual's moral agency and the intrinsic right to internal freedoms.[116] Along these lines, discussions about the emergence of 'neurorights' have developed into a rich discourse, briefly addressed in the final chapter of this dissertation (Bosoer, 2021; Inglese & Lavazza, 2021; Lavazza, 2018) (Blitz, 2010; Bosoer, 2021; Bublitz, 2013, 2015a, 2015b, 2018, 2020a; Bublitz & Merkel, 2014; Carman, 2021; Chandler, 2018; Craig, 2016; Erler, 2020, pp. 397-399; Ienca & Andorno, 2017; Inglese & Lavazza, 2021; Lavazza, 2018; Ligthart et al., 2019; Sententia, 2004; Sententia, 2013; Vallentyne, 2018c; Walsh, 2010; Wolpe, 2018; Yuste et al., 2017).

In general, in applied contexts, bioethics, public policy, and law, the discussion about neurointerventions intersects deeply with legal and ethical considerations about individual autonomy, consent, and the potential coercive use of medical technology in criminal rehabilitation, highlighting the complex interplay between medicine, law, and personal freedom. These elements shape our understanding of justice and direct the course of ethical debates. In each context, however, 'freedom' emerges as a central and paramount consideration, regardless of how we interpret its meaning. It is within this domain that the implications of neurointerventions on individual freedom and societal values become particularly salient.

Autonomy, Agency, and Mental Freedom

When we reach the realm of non-ideal theorizing in practical domains, it has been observed that notions of 'freedom' and related constituents are ubiquitous, leading to confusion in many important respects (Pugh, 2019; Pugh, 2020). This apparent contradiction refers to the mixing of ideal notions and their corresponding non-ideal instantiation in practices of policy, law, and medicine. Synthesizing the opposing domains poses significant challenges.

For example, I have discussed metaphysical notions of the 'freedom to fall.' But in an applied domain, surely such 'freedom' cannot be unfettered nor provide meaningful guidance without further explanation. It is not enough to conceive of a 'negative freedom' being limited to the absence of external constraints (Berlin, 1969). We have acknowledged the state's authority to punish, which involves restrictions on freedom. And I have argued it is valuable insofar as it facilitates the pursuit of moral virtue. But without some account of the sorts of capacities or internal mechanisms necessary to pursue this, much more must be said. I cannot explore all of these issues here, but I will highlight a few points.

'Freedom,' in the applied domains, is expressed in various forms. For example, some reference 'agency,' indicating independent action and free choices, further nuanced by 'rational agency,' suggesting reasoned, logical actions (Fischer & Ravizza, 2000; Strawson, 2010). Others argue agency also entails being able to discern the 'True and the Good' (Wolf, 1990). A more common term centres on notions of 'autonomy,' denoting self-governance, which is often used in bioethics and political theories (Feinberg, 1986) involving some kind of competency (rational decision-making, comprehension, etc.) (Frankfurt, 1988c, p. 16; Lewis, 2021, p. 16; Schaefer et al., 2014, pp. 126-127) and authenticity (alignment with core values, beliefs, and identity) (Beauchamp, 2010, p. 305; Christman, 2009, p. 102). It is further believed that factors like brainwashing, manipulation, and lack of self-awareness may compromise autonomy (Dworkin, 1988a; Friedrich et al., 2018, p. 20; Racine & Dubljevic, 2016; Schaefer et al., 2014, p. 127). So too, are discussions sometimes grounded in the notion of 'akrasia', the condition where individuals experience a disconnect between their first-order desires (immediate wants) and their second-order desires (reflective values) (Davidson, 2001b; Kalis et al., 2008; Mele, 1987).

Autonomy also connects to 'personal identity,' implying self-continuity and defining attributes, sometimes accounting for first-person perspectives (Lewis, 1976; Parfit, 1984;

Shoemaker, 1963). Autonomous individuals acting in accordance with their personal competency will be acting in alignment with their identity, although clear notions of identity may be obscure. Nonetheless, certain personal identity accounts endorse 'self-creation' (DeGrazia, 2005, p. 78) and 'narrative identity' (Baylis, 2013; Mackenzie & Walker, 2015; Schechtman, 2012). Recent 'relational' accounts consider the social aspects of concepts of autonomy, competency, authenticity, and personal identity, appreciating mind functionality and 'embeddedness', emphasizing social and normative contexts (Brown, 2015; Christman, 2004; Gallagher, 2018; Gallagher et al., 2018; Goering et al., 2021; Goering et al., 2017; Lewis, 2021, p. 19; Mackenzie & Stoljar, 2000; Meyers, 2000; Walker & Mackenzie, 2020).

This narrative traces the various themes that the applied discourse employs to strike a balance between considerations of safety and effectiveness. These considerations include whether the procedures pose risks or safety measures that are proportional to their objectives, whether their impacts are major or minor, and whether their effects are lasting or transient. At the same time, relational accounts also grapple with potential implications for human freedom by specifically addressing the immediate and foreseeable threats to alterations in personality, self-identity, and authenticity (Elliott, 1999, 2003; Glannon, 2018b; Heldke & Thomsen, 2014; Newman & Smith, 2016; Schonau et al., 2021, pp. 175-176; Tobey, 2003, p. 121).

Drawing on similar themes, there is an argument that neurointerventions may restore autonomy competencies, treat clinical pathologies, and aid in the expression of authentic identities—allowing persons to be 'active participants' in constructing or 'reconstructing' their lives (Brown, 2015; Caplan, 2006; Focquaert & Schermer, 2015, p. 147; Glannon, 2020, p. 102; Goddard, 2017; Laub, 2003, p. 281; Lavazza, 2018; Lippert-Rasmussen, 2018, p. 156; Mackenzie & Walker, 2015; Vaughan, 2006). However, the application of mandatory moral enhancements in the criminal justice system raises significant concerns. Distinguished from a consenting clinical setting, the use of neurointerventions for criminal behaviour incites a divided debate about the practical and conceptual possibility of improving or 'enhancing' moral capacities (Buchanan, 2011; Carman, 2021; Conan, 2020; Donnelly-Lazarov, 2021; Earp et al., 2018; Focquaert & Schermer, 2015, p. 139; Harris, 2011, 2014a; Huang, 2020; Levy, 2020; Lewis, 2021; Palk, 2018; Persson & Savulescu, 2008, 2011a, 2011b, 2012, 2013, 2015;

Rüther & Heinrichs, 2019; Savulescu & Persson, 2012; Sparrow, 2013; R. Sparrow, 2014; R. J. Sparrow, 2014; Wiseman, 2016).

In the context of quasi-coercive measures, individuals may come to accept neurointerventions as beneficial, potentially reconciling the tension between first-order desires (immediate wants) and second-order desires (long-term reflective values) through a process akin to self-filtering, which can be seen as an expression of enhanced freedom rather than restriction. This highlights a nuanced dimension of freedom, where the alignment of desires through neurointerventions might allow individuals to achieve a higher expression of their authentic selves and rational agency.[117]

This underscores one potential response to the 'freedom objection', which posits that in certain situations, neurointerventions may indeed enhance human freedom—depending on how we conceive it. This complex issue merits further exploration. A framework that discerns where neurointerventions may boost freedom is highly sought after. For example, I have supported the 'therapeutic justice movement,' which employs consent-based neurointerventions, often alongside traditional therapies, to address substance use disorders that I believe can be highly destructive to the capacity for autonomous human agency (Coppola, 2018, p. 9; Hardcastle, 2020, p. 162; Marlowe, 2021; Matravers, 2018, p. 72; Matusow et al., 2013; Sifferd, 2020, p. 309; Tonry, 2011b).

While I advocate for the use of this technology, I posit that in the vast majority of cases for competent persons, it must involve the individual's consent and participation. This allows for the self-driven choice to undertake operations for self-improvement and the subsequent integration of changes, preserving autonomy and feelings of self-authorship. Such an approach mitigates concerns about agency, authenticity, and identity.

However, as previously contended, neurointerventions, especially in cases of involuntary interventions or broad-scale moral modifications, present non-trivial challenges. These encompass distinguishing genuine capacity enhancement from mere behavioural control, respecting individual predispositions and innate characteristics, wrestling with the absence of a consensus on appropriate moral judgement models, translating moral theories into complex real-world situations, and recognizing the importance of embedded and extended agency (Banja, 2018; Batson et al., 2009; Carman, 2021, p. 192; Deonna & Teroni, 2012; Dubljevic & Racine, 2014; 2017, pp. 340-342; Earp et al., 2017; Earp et al., 2018, p. 167; Focquaert &

Schermer, 2015; Greene et al., 2008; J. D. Greene et al., 2001; Haidt & Graham, 2007; Harris, 2013b; Lewis, 2021, p. 18; McMillan, 2014; Mikhail, 2011; Persson & Savulescu, 2012; Raus et al., 2014; Shook, 2012, p. 5; Wiseman, 2016; Wudarczyk et al., 2013).

At this stage, I think it is sufficient to identify that the permissibility of neurointerventions at an applied level turns on considering various aspects of freedom alongside many diverse concepts. In the domain of biomedical ethics, public policy, and law, notions of 'freedom' ultimately guide action; however, ideals used in non-ideal circumstances involve complexities which demand far more nuanced understandings of freedom. At the very least, these domains reveal the deeply ingrained assumptions of 'freedom' that guide our activities and atop which we build many guides for action. These ingrained norms, however, do allow us to perceive the normative asymmetries that arise with imposed neurointerventions.

On that note, a transition is available to the parallel and intertwined idea of the 'rights.' I have previously argued for the recognition of a right to 'mental self-determination' or 'mental integrity' and that it is conceivable such a right could be grounded in exploring certain of these features: agency and interrelated concepts such as personal identity, competency, authenticity, and more broadly, self-authorship (Craig, 2016, p. 112). In this sense, a right to mental integrity might be defended, on some accounts, as a necessary corollary to 'mental freedom.'[118] As I will briefly show in the concluding remarks of this thesis, many of the norms guiding the applied domains ultimately involve or would benefit from careful scrutiny of the rights discourse.

Conclusion—Freedom, Equivalence, and Asymmetry

In this chapter, I have embarked on an exploration of punishment equivalence and the moral and normative asymmetry surrounding neurointerventions in the criminal justice system. The vast scope and complexity of this topic necessitate ongoing inquiry and a deepened understanding. While it is impossible to address all intricacies within each domain, this exploration has identified pressing issues that warrant further examination.

Engaging with ideal and non-ideal theories, as well as drawing insights from various disciplines, I have aimed to shed light on the ethical complexities that arise when considering neurointerventions as alternatives to traditional forms of punishment. By doing so, I have established a common theme: the plausibility of moral and normative asymmetry between the use of neurointerventions and conventional forms of punishment based on respective threats to

‘mental freedom.’ It is possible these asymmetries have profound implications for moral agency, human dignity, and the intricate interplay within the community of minds.

However, it is crucial to acknowledge the limitations of our current understanding and the need for ongoing research, discourse, and responsible theorizing to address the ethical complexities associated with neurointerventions. The exploration conducted in this chapter serves as a starting point, offering important insights that pave the way for further investigation.

But as we descend from where we began this chapter—the realm of lofty ideals—to where we concluded—the circumstances of the real world, we are reminded of the ethical pitfalls that demand our attention. From the haunting echoes of a dark history of prisoner rights violations to the cautionary tales of lobotomy and human experimentation, we must confront these complexities with humility and a profound sense of responsibility.

Whether through nonconsensual neurointerventions or the labyrinth of depraved prison conditions and alive to lessons from history, there is one equivalence I think bears note. Both involve the use of state power and the risk of committing serious harm. Safeguarding the dignity, agency, and well-being of all individuals requires a path guided by justice, compassion, and unwavering responsibility.

Concluding Remarks

Mind, Rights, and the Path Ahead

Retracing the Path—Rhetorical Musing

This journey set out to address a seemingly simple query, '*If locking criminals up in prison is justified, why not require them to undergo some type of 'safe and effective' neurointervention*?' But this exercise has led us, I think, to deeper intricacies of the individual psyche and into the vast realms of a broader societal and normative landscape. The realms of freedom, rationality, and darker aspects of our nature while reflecting on the recursive relations between the individual mind and the broader social fabric. In the context of criminal offenders, those we would rather forget, perhaps observing a bridge in the gap between 'us' and 'them'—as we come to see ourselves for who we really are.

So it seems we near the end of our voyage. We have trudged through penal theory, scientific jargon, technical philosophical nomenclature, normative schematics, and metaphysical doctrine. And under the spectre of a 'post-human future,' I recount this strange journey with a touch of flourish, perhaps unbefitting academic discourse, but I think, fittingly accentuating its very humanness.

Punishment

Our journey began in the depths of a dark place. One where perceived transgressors, such as young Capay, at this very time, remain carefully stowed away. But also a dark place deep in ourselves, reflecting relics of a primitive time long past and vestiges of a myopic and vengeful nature ill-suited to our lofty aspirations. This is a hard truth I have argued we must face before we can begin to address the crisis in our criminal justice practices, which continues to cause untold suffering for many who deserve humanity and compassion.

For the 'punishment claim,' I maintain *in theory*, we can justify punishment. However, in practice, in our institutions of punishment, and in the real world, we cannot. So, if equivalency arguments claim to use our current practices of punishment as a 'baseline' for comparison of the relative threats of neurointerventions, they face significant and perhaps insurmountable

challenges. Equivalency arguments cannot offer practical, meaningful guidance for real-world implementation in a substantial range of instances.

But I also suggested this compels us to question from the outset whether grounding discussions about 'treating crime' through forcible intervention in the brain of the disfavourable and vulnerable is misguided from the outset. Perhaps the most promising solution is to use the powerful tools contemporary neuroscience offers to first reproach within ourselves the faults we are quick to see in the 'deviant other'—and were it possible, I believe, likely our best hope for building a society that is less brutal, less fearful, and more humane.

Brain, Mind, Safety, and Efficacy

We then continued with an exploration of a realm—perhaps best left to those hard at work in laboratories and well-versed in affective and cognitive neuroscience—that nonetheless served the purpose of endearing both wonder and reticence. A three-pound organic mass composed of chaotic particles from a universe that has somehow become conscious and aware of itself. And appreciating throughout our lifetimes, through a broader social and normative discourse, shared with those present and those that have passed, arrays and configurations and possibilities that exponentially exceed the particles in the known universe from which it has come to be assembled. And all the while, at the very moment we come to ponder such a curiosity, it is that same curiosity that facilitates the very exercise we undertake—to construct meaning out of an otherwise chaotic flow of information.

Faced with such a marvel and owning up to our past failures, I argued we must take care not to cause harm from a gross overstep. Armed with powerful tools for neuromodulation and laying bare our assumptions, we should pause with humility, acknowledging the gulf between what we know and what we could perhaps hope to know.

For the In-Principle Claim, I argued the vast complexity of the brain poses substantial challenges, both practical and theoretical. The difficulty lies in accurately identifying and safely adjusting areas related to moral judgment using neurotechnologies due to the brain's intricate nature. Moreover, even if we could perfect such technology, we must still establish a standard for moral judgment. These judgments reach beyond brain activity, involving our interpretations of ethics, duties, and responsibilities and a larger normative discourse into which the mind itself extends. Hence, efforts to find answers in neuroscience must also include a thorough

understanding of the wider context that shapes our moral world. This, in turn, might open the door to understanding how to conceive freedom in ways that are of interest to ethical discussions.

The Principle of Parity

After trudging through the complex realm of neuroscience, we continued to consider the strange case of what, at first sight, seemed a fairly odd man running a half marathon barefoot above the Arctic Circle before shifting to monks meditating in peaceful serenity—both eventually hooked up to brain scanners, in an attempt to identify matters of interest. We dug a little bit further into the deeper recesses of our primitive minds, and if the exercise was successful, we were able to learn a little more about the limits of our rationality—the source of those 'vapours, odd beings, terrors, and deluding images' which cloud our judgement, and obscure the higher aspects of our nature. The same that leaves us susceptible to distortions and powerful influences that confound us on every front. But I think, on further reflection, what was found was at least a flicker of hope.

For the parity principle, I hoped to show that 'ideal theorizing' has not accounted for a significant facet of our mental life, from which springs various avenues for theorizing. Specifically, emergent properties of the mind, such as its ability for self-organization—among other things, through a form of 'internal mentation'—the temporal aspects of mental states and the way the mind extends into a broader social and normative discourse. In quiet moments of introspection, perhaps, we find refuge. Reflecting on internal mentation, our brain exhibits a remarkable ability to self-organize and resist external stimuli, even in moments devoid of any interventions that cannot be attributed solely to environmental stimuli, indicating an inherent capacity for resistance. This inner sphere defies simple explanation by 'direct or indirect' interventions and potentially suggests a wider sphere for individual liberty. That we are not mere creatures of happenstance—or at least, might aspire not to be.

And while it is almost certainly true that the limits of our rationality and widespread indirect subversion are far more powerful influences at present than the sorts of neurointerventions available, I suggested, at least in principle, this need not be the case. As so too, we cannot rule out the possibility that neurointerventions may pose a unique, and distinct threat. To echo the prose from our fictional starting point with the *Shawshank Redemptions*: "Fear can hold you prisoner. Hope can set you free" (Darabont & King, 1994a).

Equivalence and Asymmetry

In the end, we turned to consider what came out of this exercise. Perhaps equipped with a better understanding of ourselves, we considered where we might find a balance between unfettered human freedom and the 'hidden demons of our nature.' Entertaining the possibility of moral failures or 'falls' can be transformative, leading to personal growth, increased empathy, and a more profound understanding of morality—a possible path to *eudaimonia* or 'human flourishing.' And we considered how such a path might be possible under the watchful eye of the 'God Machine,' a revived deity of our own creation seeking to proscribe transgression.

Those reflections, I suggest, underscore the potential value and delicacy of human nature, which, within appropriate boundaries, merits safeguarding. And we traced it through various spheres of theorizing that cast light on the questions about what we might find 'right or wrong.' The exercise served, if successful, to highlight a need to explore the significance of 'mental freedom,' which finds justification in themes traced across various domains of analysis in contemporary moral, ethical, political, and legal theorizing. Not just to preserve individual identity and autonomous decision-making but also to maintain the collective ethical and moral fabric of society from which this very discourse springs.

I argued, in the end, we could not be confident neurointerventions and conventional forms of punishment are *equivalent* or rule out the risk; even at the highest level of theoretical analysis, they pose unique risks or present unique 'wrongs' however we conceive of the term. And that all coalesce around an elusive but imitable notion of 'mental freedom.'

In technical terms, there was a *moral* asymmetry because neurointerventions pose distinct threats to human rationality and mental freedom, which generates moral reasons against implementation that did not hold for conventional forms of punishment. There was a *normative* asymmetry because, alongside these moral reasons, there were distinct practical and prudential reasons against using neurointerventions—particularly given real-world circumstances.

In the end, I hoped to establish that the question: *If locking criminals up in prison is justified, why not require them to undergo some type of 'safe and effective' neurointervention*?' affords many answers, but at best, there are very serious reasons to exercise extreme caution before beginning discussions about a system for practical implementation in the non-ideal circumstances of the world that we live in. Even at the highest level of ideal theorizing, it is far

from certain that we can rule out the possibility of a morally relevant difference. That difference being 'freedom,' or perhaps 'mental freedom'—however, we do, or might come to understand it.

And so, I argued, it is our responsibility to navigate these issues conscientiously, preserving humanity amidst rapid advancements in neurointerventions and the endangerment of the person. But acknowledging far more theorizing is required, at least, in addressing questions about neurointerventions and crime prevention, we might be satisfied that in the end, perhaps we are met with fewer answers than questions—and far fewer answers to the further questions raised. In closing, I briefly address one of these questions.

A Path Forward: The Nascent Right to Mental Integrity?[119]

I would conclude with a very brief mention of a question that I cannot hope to fully explore here. Instead, I flag it as somewhat of an 'epilogue' for future discussions. During early research, the topic of interest I hoped to consider was framed by a different question and intuition.

There is a strong intuition there is something valuable, inimitable, and fragile about the mind and its primary carrier, the brain. As Neil Levy states: "If we have the right to a sphere of liberty, within which we are entitled to do as we choose, *our minds* must be included within that sphere" (Levy, 2007, p. 179). And if this is the case, as I have suggested, it is possible up to this point. The question is *how* we might extend protection to it. As I think we have seen when it comes to issues of the mind and neurointerventions, what begin as simple questions seldom remain so. And so I frame a further question: *do persons have rights over their own minds*? Beginning to address this question would, I have come to see, require an entire dissertation and perhaps an entire career in its own right. I simply highlight a few issues here.

Mental Rights—Intersections

I have argued we cannot rule out the risk there was a *moral asymmetry* between direct and indirect interventions—comparing neurointerventions and conventional forms of punishment, such as incarceration. I argued the latter could, *in principle*, pose a greater threat to 'human freedom,' as we might understand it across various domains. I suggested in response to the distinct threats of neurointerventions, we might conceive of this as a form of 'mental freedom'—reflected in the title of this dissertation—also leading to a *normative asymmetry* at

the level of non-ideal theorizing. But if this is the case, do we have moral or prudential reasons to extend protection, and if so, how would this be realized?

One theme we have traced throughout this work, and that has arisen at various junctures, is the concept of *rights*. In Chapter 2, we broached 'rights forfeiture' as a punishment justification. This theory suggests that crime commission implies certain rights forfeiture, thus 'lowering moral barriers'—which I suggested required accounting for a broader, more nuanced structure of 'contoured moral rights.'

I argued that this presented challenges for equivalency arguments because the reasons and justification the state has for justifying punishment do not necessarily account for the countervailing considerations that might restrict that power—which required accounting for rights. In the final chapter, I argued that the sorts of normative considerations that arise around the use of neurointerventions in criminal justice practices would ultimately involve implementation in the real world, and practically speaking; this would require them to be considered in the context of contemporary social practices that include rights.

In previous work, I have explored research discussing a proposed right to 'mental integrity' that would protect such core features—rational agency, personal identity, and authenticity—all grounded in some general notion of 'self-authorship' (Craig, 2016). (Craig, 2016). And it has been explored by others in depth over the past decade across various domains and fields of study, tracing themes about rights to 'freedom of thought' and *nascent* rights to 'mental self-determination' (Blitz, 2010; Bosoer, 2021; Bublitz, 2013, 2015a, 2015b, 2018, 2020a; Bublitz & Merkel, 2014; Carman, 2021; Chandler, 2018; Craig, 2016; Erler, 2020, pp. 397-399; Ienca & Andorno, 2017; Inglese & Lavazza, 2021; Lavazza, 2018; Ligthart et al., 2019; Sententia, 2004; Sententia, 2013; Vallentyne, 2018c; Walsh, 2010; Wolpe, 2018; Yuste et al., 2017).

However, the existence and nature of a 'right to mental integrity' is a complex issue. As we have seen throughout this dissertation, our minds and brains are subject to constant influence and manipulation in numerous ways, many of which we willingly accept or even find beneficial. So, the notion of an 'inviolable' or 'unqualified' right is surely not plausible.

Even in ideal cases, there are limits to our rationality. In others, the realization of any lofty aspiration of 'freedom' would still require more than negative freedom; instead, perhaps

positive rights or one that identifies the potential benefits of neurointerventions to restore competencies necessary for the effective realization of freedom.

Also, as I have argued, the brain and mind are not the same—the latter is a much broader thing. I have argued the same holds for human rationality and 'mental freedom.' Taken to the extreme, the latter could be seen to extend throughout vast expanses, making any proposed right over the 'mind' something that might, itself, be incomprehensible—falling somewhere between 'everything' and 'nothing.' What would such a right hope to protect? These are significant challenges that would need to be addressed.

Even leaving all this aside, the problem, as I see it, moving forward, is the complexity of rights. Much like the mind, the discourse of rights itself is rich and complex, comprising social practices, interpersonal and state relations, and a nuanced matrix of moral, political, and legal accounts, asymmetrically fragmented across various domains and viewed through diverse lenses (Campbell, 2017; Devlin, 2009; Dworkin, 1977, 1985, 1986; Feinberg, 1978; Fuller, 1958; Green & Adams, 2019; Hart, 1955, 1961 [2012], 1963, 1979; Hart, 1958; Pardo, 2014; Plunkett & Shapiro, 2017; Rawls, 1955; Rumble, 1832 [1995], p. 157; Tadros, 2011).

Navigating this web involves taking into account various conceptions of rights—be them moral, political, or legal—with varying source, nature, and normative grounding (Dworkin, 1986; Dworkin & Waldron, 2013; Feinberg & Narveson, 1970; Flathman, 1976; Gardner, 1997, 2004; Goldman, 1979; Habermas, 1996a; Hart, 1961 [2012]; Hohfeld, 1913; Ivison, 2007; MacIntyre, 1984, p. 187; Minow, 1990; Nedelsky, 1993, p. 153; Nozick, 1974, pp. 28-33; Raz, 1984, 1999; Scanlon, 2000, 2014, 2020; Schroeder, 2007; Steiner, 2006, p. 470; Steyn, 2004; Taylor, 1973, pp. 32-34; Wellman, 1987; Wellman, 2020; Yowell, 2007). Among other things, justifying rights involves the exploration of moral or normative assertions about morality and value, encompassing recognition, respect, dignity, well-being, equality, fairness, autonomy, and positive freedom (Beitz, 2015; Gewirth, 1978, 1985, 1998; Gilabert, 2019; T. H. Green, 1986; Green, [1883] 1990; Griffin, 2009; Hegel, 1987, 2017 [1807], [1821] 1991; Hobbes, [1651] 1980; Locke, 1824c, [1698] 1998; Neuhouser & Neuhouser, 2009; Nickel, 1987; Nussbaum, 1992, 2009; Patten, 1999; Rawls, 2001, pp. 42-50; Sen, 2001; Sumner, 1987; Talbott, 2010; Tasioulas, 2015; Taylor, 2015). And as we have seen, such terms are deeply problematic when we claim to ascend anywhere into the realm of first-order moral theorizing.

Moreover, as societal and technological environments continue to evolve, so too does the landscape of rights. This gives rise to longstanding issues about *what rights there are*, concerns about conflicting rights and priorities and the criteria for recognizing rights, further issues about the procedure for recognizing new rights, how specific they are, and the need to avoid a 'rights explosion' (Brems, 2009; Rawls, 1999; Raz, 2007). Any development of 'rights over the mind'—or perhaps, specific rights against neurointerventions—would need to face all of these challenges and more.

Further, we face the challenge of punishment—any such rights would need to be realized in social practices that recognize punishment often requires the restriction of rights—and various accounts of how, why, and to what extent this could occur (Goldman, 1979; Matravers, 2018; Morris, 1991, p. 68; Quinn, 1985, pp. 332-333; Simmons, 1994, p. 149; Wellman, 2009, 2012, 2017, 2020; Gardner, 1997, 2004; Goldman, 1979; Lippke, 2001, pp. 85-87; Martin, 1993; Quinn, 1985; Raz, 1999; Simmons, 1994, p. 158; Wellman, 2020).

A survey into the discourse on rights thus poses significant obstacles to both addressing these raised issues and grounding productive discussion going forth. Yet, as technology advances and neurointerventions become more pressing, there appears to be a demand for the literature on rights to evolve in order to protect those most central freedoms of the human condition.

The Discourse of Human Rights

This aside, I think the discourse on human rights presents an important and rich area of study for future research. This research is rooted in political, legal, and social structures in both ideal and applied domains. Within this research is the central idea of 'Human Rights,' often viewed as 'special rights,' which are seen to play a unique role in the 'global basic structure' due to their reactive nature and the ability to protect critical individual interests against foreseeable threats (Beitz, 2011). In relation to human rights, two compelling expansionary eras—the Enlightenment and post-WWII—stimulate this rich discourse (Beitz, 2001, 2011; Buchanan, 2007b; Edmundson, 2012 part II; Morsink, 2009; Neier, 2013; Pogge, 2005; Tuck, 1979).

The former represented a time of intellectual and philosophical awakening, leading to the codification of inalienable human rights in response to the period's systemic oppression and social inequities. This evolution set the foundation for understanding individual liberty and freedom from undue intrusion (Hunt, 2007; Ishay, 2004). It also arose alongside a belief in the

power of human reason to overcome nature and significant advancements in the sciences and medical sciences to improve human health and well-being—culminating in advanced contemporary technologies, such as neurointerventions (Bronstein, 2010, p. 85).

The latter expansionary era of international human rights arose in response to horrific human rights violations, including non-consensual human experimentation using novel biomedical technologies perpetrated during WWII. The aftermath of these atrocities led to the formation of the Universal Declaration of Human Rights (UDHR) in 1948 by the United Nations. The Preamble of the UDHR emphasizes the 'recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family,' which is the 'foundation of freedom, justice and peace in the world' (Glendon, 2002; Ishay, 2008; Morsink, 1999). It also served as the groundwork for subsequent international human rights treaties, including the International Covenant on Civil and Political Rights, which explicitly prohibits 'torture or cruel, inhuman or degrading treatment or punishment' and 'medical or scientific experimentation without free consent' (Article 7). This came as a direct response to the atrocities committed during WWII and solidified the commitment of the international community to prevent such violations.

One could also argue that the discourse on human rights—rooted in political, legal, and social frameworks—provides valuable avenues for discussion and theorizing on the ethical use of neurointerventions in practical domains. In this sense, the reactionary nature of human rights, grounded in deontological sentiments, notions of respect, and human features or virtues, may show a promising area of inquiry. It reminds us that any implementation of such technologies must respect the inherent dignity and inalienable rights of individuals, adhering to principles of consent, freedom, and justice.

However, the realm of human rights is characterized by inherent pluralism and is subject to various critiques and divisive debates. As we have seen through our exploration in Chapter 4, there is a real risk that general references to 'dignity,' 'human form,' and 'justice' are far from uncontentious, and responsible philosophical grounding would be required to render them more than mere hyperbole in addressing novel technologies. Nonetheless, I believe tracing some of those themes in Chapter 4 has shown connections and certain conceptual groundings that show promise.

Dependence, Interdependence, and Indivisibility

As a final point of interest, if we accept the existence of rights over the brain or mind—through the expansion of current rights or recognition of new ones—it is worth exploring whether these might warrant special protection against nonconsensual neurointerventions. This coalesces around the architecture and connection between rights.

In contemporary rights theory, the concept of 'supporting rights,' 'indivisibility,' and 'interdependence' explores the complex relationships between different rights, emphasizing 'basic' or 'fundamental' rights. Henry Shue advocates for recognizing rights such as security and subsistence as 'basic' due to their critical role in implementing other rights (Kuflik, 1984; Shue, 2020). It serves to highlight a unique architecture and hierarchy of rights and establishes the foundational principles upon which more elaborate theories are built (Elster, 2015; Gilabert, 2010; Kuflik, 1984; Nickel, 2008, 2010, 2016; Philips, 2014; Quintavalla & Heine, 2019; Sen, 2001).

James Nickel's work significantly contributes to this discourse. He introduces a complex theory of 'supporting rights' that employs concepts like 'dependence,' 'strong support,' and 'indispensability' to define the relationships between different rights. Nickel presents a compelling argument that endorsing one right is often logically or practically inconsistent without the simultaneous implementation of another right. His analysis delves into the nature of these relationships, exploring both unidirectional and bidirectional dynamics, and introduces the concept of 'system-wide relations' (Nickel, 2008, 2010, 2016).

For example, Nickel posits that two rights are indivisible when there exists strong bidirectional support between them—meaning the full realization of one right cannot occur without the full realization of the other. A prime example Nickel presents is the relationship between the right to free and regular elections and the right to political participation. This strong bidirectional support is indicative of the intricate interplay and mutual reliance existing within the system of rights (Nickel, 2008, p. 990).

I think, in future theorizing, it might be useful to consider how these theories might apply in the context of the use of neurointerventions—particularly as part of criminal justice practices—and the manner in which rights over the mind might fit into discussions about legal protection. I have previously argued the brain or 'mind' is intimately tied to rational agency, autonomy, personal identity, and 'self-authorship'. Arguably, these rights may deserve a special place in the

architecture of rights, given their role in enabling our comprehension and exercise of many, if not most, other rights.

In a previous article, I addressed this issue with reference to the hypothetical example of the 'Damascus Drug,' which permanently changed a violent psychopath (Saul) into a religious pilgrim (Paul) (Craig, 2016, pp. 112-113). As I observed:

> The drug has a non-trivial effect on core features associated with Saul's agency, specifically his personal identity, competency, authenticity, and general capacity for self-authorship. Further, by forcibly administering the drug, the state violates Saul's bodily integrity. But it also implicates a broader spectrum of his rights—his freedom of religion, movement, association, expression, conscience, and his right to free and full development of personality (Craig, 2016, p. 113)

The thought experiment elicited diverse responses, underscoring the complexity of defining such rights, the requirement for a balance with social objectives, such as punishment, and parity considerations (Petersen & Kragh, 2017; Ryberg, 2020, pp. 70, 84, 91-92)—all pressing issues I acknowledge.

But recognizing the vital role of brain-related rights in understanding and exercising all other rights fosters reflection. I maintain my belief in their importance in illustrating the stakes, particularly when acknowledging unique threats to human agency, questioning the parity principle, and distinguishing moral from normative asymmetry. A constant motif in this exploration is the interconnectedness among individuals and wider normative discourse. This dynamic provokes analogous questions about the place of mind-related rights within it and emphasizes the inherent 'humaneness' of such an inquiry, hoping to inspire further explorations in this stimulating terrain.

Arriving Where We Began and the Path that Lies Ahead

In conclusion, the author admits traversing the intricate realm of rights has been both illuminating and humbling. It provided a fresh perspective into the intricate web of moral, political, and legal discussions that shape our perception of human rationality and freedom and, perhaps, inform our path forward. The true value of this journey is in recognizing the vast expanse

of intellectual territory yet unexplored. Thus, we return to our starting point with a newfound understanding, recalling Eliot's words, 'We know the place for the first time' (Eliot, 1943).

Through neurotechnology, freedom, and autonomy, we have embarked on a profound examination of our humanity, advocating for externalist ethics, as Levy suggests (Levy, 2017; Levy, 2019; 2020, pp. 45-46), acknowledging our intertwined existence with our environment. Recognizing our unique nuances and societal interconnectedness, we aim to maintain humility, cognizant of the expansive unknown that lies ahead. As our understanding continues to evolve, humility keeps us grounded and respectful of our human essence, which warrants careful reflection and potential protection.

As this journey draws to a close, it is not, I think, with a final period but rather with a gentle bookmark, pausing an ongoing exploration. A sense of disappointment for unanswered questions is eclipsed by unending curiosity and the allure of yet unasked inquiries. This conclusion heralds the dawn of many future endeavours, each promising further exploration into the unexplored terrains of our knowledge.

As we confront challenging questions about whether or not we should wield the power of the state to forcibly intrude into the brains of transgressors, including not only callous wrongdoers but the disadvantaged and despised, there is a need for caution. I have argued that it is perhaps our capacity to recognize the limits of our rationality that provides us with the most powerful tools to overcome those limits. And so, I end with a note of caution as we navigate intricate ethical and legal challenges, especially those involving unfavourable individuals, realizing the profound implications on legal institutions, social structure, and society.

And as we grapple with these issues, caution is warranted. Our ability to recognize our rationality's limits and, in turn, become more than creatures driven by powerful subversion and mere creatures of happenstance—that whatever lies ahead is not entirely outside of our control. Alive to this risk, we should heed the wisdom of Viktor Frankl—derived from his experiences in concentration camps, those living laboratories and testing grounds. That our potentials are actualized based on decisions and not conditions, and to learn to see ourselves as we truly are: "After all, man is that being who invented the gas chambers of Auschwitz; however, he is also that being who entered those gas chambers upright, with the Lord's prayer or the Shema Yisrael on his lips" (Frankl, 1985, p. 134).

Bibliography

Abbott, M. N., & Peck, S. L. (2016). Emerging Ethical Issues Related to the Use of Brain-Computer Interfaces for Patients with Total Locked-in Syndrome. *Neuroethics*, *10*(2), 235-242. https://doi.org/10.1007/s12152-016-9296-1

Adamczyk, A. K., & Zawadzki, P. (2020). The Memory-Modifying Potential of Optogenetics and the Need for Neuroethics. *NanoEthics*, *14*(3), 207-225. https://doi.org/10.1007/s11569-020-00377-1

Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, *14*(1), 43-64.

Adams, R. M. (2006). *A theory of virtue : excellence in being for the good*. Clarendon.

Adebowale, V. (2010). Diversion, not detention. *Public Policy Research*, *17*(2), 71-74. https://doi.org/10.1111/j.1744-540X.2010.00607.x

Altimus, C. M., Marlin, B. J., Charalambakis, N. E., Colón-Rodriquez, A., Glover, E. J., Izbicki, P., Johnson, A., Lourenco, M. V., Makinson, R. A., McQuail, J., Obeso, I., Padilla-Coreano, N., & Wells, M. F. (2020). The Next 50 Years of Neuroscience. *The Journal of Neuroscience*, *40*(1), 101-106. https://doi.org/10.1523/JNEUROSCI.0744-19.2019

Altimus, C. M., Marlin, B. J., Charalambakis, N. E., Colon-Rodriquez, A., Glover, E. J., Izbicki, P., Johnson, A., Lourenco, M. V., Makinson, R. A., McQuail, J., Obeso, I., Padilla-Coreano, N., Wells, M. F., & for Training Advisory, C. (2020). The Next 50 Years of Neuroscience. *J NEUROSCI*, *40*(1), 101-106. https://doi.org/10.1523/JNEUROSCI.0744-19.2019

Alvarez, M. (2009). Actions, thought-experiments and the 'Principle of alternate possibilities'. *Australasian Journal of Philosophy*, *87*(1), 61-81. https://doi.org/10.1080/00048400802215505

Anderson, J. K. (2010). The Decentralization of Morality in "Paradise Lost". *Rocky Mountain review (Rocky Mountain Modern Language Association)*, *64*(2), 198-204.

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human

prefrontal cortex. *NAT NEUROSCI*, *2*(11), 1032-1037. https://doi.org/10.1038/14833

Andrews-Hanna, J. R. (2012). The brain's default network and its adaptive role in internal mentation. *Neuroscientist*, *18*(3), 251-270. https://doi.org/10.1177/1073858411403316

Andrews, K. (2020). Naïve Normativity: The Social Foundation of Moral Cognition. *Journal of the American Philosophical Association*, *6*(1), 36-56. https://doi.org/10.1017/apa.2019.30

Annas, J. (2003). Virtue ethics and social psychology.

Annas, J. (2011). *Intelligent virtue*. Oxford University Press.

Ansari, D. (2012). Culture and education: new frontiers in brain plasticity. *TRENDS COGN SCI*, *16*(2), 93-95. https://doi.org/10.1016/j.tics.2011.11.016

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy*, *33*(124), 1-19.

Arboleda-Florez, J. (2005). The ethics of biomedical research on prisoners. *Curr Opin Psychiatry*, *18*(5), 514-517. https://doi.org/10.1097/01.yco.0000179489.70014.d3

Ariely, D. (2010). Predictably irrational: the hidden forces that shape our decisions. *Math Comput Educ*, *44*(1), 68.

Aristotle. (2009). *Nicomachean Ethics* (T. Irwin, Trans.; L. Brown, Ed.). Oxford University Press.

Arpaly, N. (2002). *Unprincipled virtue: An inquiry into moral agency*. Oxford University Press.

Arstila, V. (2014). *Subjective time : the philosophy, psychology, and neuroscience of temporality*. MIT Press.

Atmanspacher, H. (2020). Quantum approaches to consciousness. In E. N. Zalta (Ed.): Standford Encyclopedia of Philosophy.

Ayer, A. J. (1971). *Language, truth and logic*. Harmondsworth, England : Penguin Books.

Aylett, M., Mahmut, M., Langdon, R., & Green, M. (2006). Social cognition in nonforensic psychopathy: further evidence for a dissociation between intact 'theory of mind'and impaired emotion processing. *Acta neuropsychiatrica*, *18*(6), 328-329.

Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

Baars, B. J. (1997). *In the theater of consciousness: The workspace of the mind*. Oxford University Press, USA.

Baggio, G. (2018). *Meaning in the brain*. Cambridge, Massachusetts : The MIT Press.

Bahr, S. J., Masters, A. L., & Taylor, B. M. (2012). What works in substance abuse treatment programs for offenders? *The Prison Journal*, *92*(2), 155-174.

Bales, W. D., & Piquero, A. R. (2012). Assessing the impact of imprisonment on recidivism. *Journal of Experimental Criminology*, *8*(1), 71-101.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Pers Soc Psychol Rev*, *3*(3), 193-209. https://doi.org/10.1207/s15327957pspr0303_3

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of personality and social psychology*, *71*(2), 364.

Banja, J. D. (2018). Moral Reasoning. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: direct effects of trait construct and stereotype-activation on action. *J Pers Soc Psychol*, *71*(2), 230-244. https://doi.org/10.1037//0022-3514.71.2.230

Barker, V. (2017). Nordic vagabonds: The Roma and the logic of benevolent violence in the Swedish welfare state. *European journal of criminology*, *14*(1), 120-139. https://doi.org/10.1177/1477370816640141

Barn, G. (2019). Can Medical Interventions Serve as 'Criminal Rehabilitation'? *Neuroethics*, *12*(1), 85-96. https://doi.org/10.1007/s12152-016-9264-9

Baskin-Sommers, A. R., & Fonteneau, K. (2016). Correctional change through neuroscience. *FORDHAM LAW REV*, *85*(2), 423-436.

Baskin, J. H., Edersheim, J. G., & Price, B. H. (2007). Is a picture worth a thousand words? Neuroimaging in the courtroom. *Am J Law Med*, *33*(2-3), 239-269. https://doi.org/10.1177/009885880703300205

Bassett, D. S., & Gazzaniga, M. S. (2011). Understanding complexity in the human brain. *TRENDS COGN SCI*, *15*(5), 200-209. https://doi.org/10.1016/j.tics.2011.03.006

Bassett, D. S., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, *20*(3), 353-364.

Bastian, B., Denson, T. F., & Haslam, N. (2013). The roles of dehumanization and moral outrage in retributive justice. *PLOS ONE*, *8*(4), e61842. https://doi.org/10.1371/journal.pone.0061842

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *BIOL LETT*, *2*(3), 412-414. https://doi.org/10.1098/rsbl.2006.0509

Batson, C. D., Chao, M. C., & Givens, J. M. (2009). Pursuing moral outrage: Anger at torture. *Journal of Experimental Social Psychology*, *45*(1), 155-160. https://doi.org/10.1016/j.jesp.2008.07.017

Battro, A. M., Fischer, K. W., & Léna, P. J. (2008). *The Educated Brain: Essays in Neuroeducation*. Cambridge u.a: Cambridge University Press. https://doi.org/10.1017/CBO9780511489907

Baylis, F. (2013). "I am who I am": on the perceived threats to personal identity from deep brain stimulation. *Neuroethics*, *6*(3), 513-526.

Baylis, F. (2019). *Altered inheritance : CRISPR and the ethics of human genome editing*. Cambridge, MA : Harvard University Press.

Bear, M., Connors, B., & Paradiso, M. A. (2020). *Neuroscience: Exploring the brain*. Jones & Bartlett Learning, LLC.

Beauchamp, T. L. (2010). Autonomy and consent. *The ethics of consent: Theory and practice*, 55-78.

Beauchamp, T. L. (2013). *Principles of biomedical ethics* (7 ed.). Oxford University Press.

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7th ed.). Oxford University Press.

Bechtel, W. (2007). *Mental Mechanisms*. London: Psychology Press. https://doi.org/10.4324/9780203810095

Bedau, H. A. (1972). Penal Theory and Prison Reality Today. *Juris Doctor*, *2*, 40-43.

Bedau, H. A. (1978). Retribution and the Theory of Punishment. *The Journal of philosophy*, *75*(11), 601-620.

Bedau, H. A., & Kelly, E. (2019). Punishment. In E. N. Zalta (Ed.), (Winter 2019 ed.). Metaphysics Research Lab, Stanford University. url{https://plato.stanford.edu/archives/win2019/entries/punishment

Beeker, T., Schlaepfer, T. E., & Coenen, V. A. (2017). Autonomy in Depressive Patients Undergoing DBS-Treatment: Informed Consent, Freedom of Will and DBS' Potential to Restore It [Hypothesis and Theory]. *Front Integr Neurosci*, *11*(11), 11. https://doi.org/10.3389/fnint.2017.00011

Beetham, D. (2013). *The legitimation of power*. Bloomsbury Publishing.

Behme, C. (2013). Assessing Direct and Indirect Evidence in Linguistic Research. *Topoi*, *33*(2), 373-383. https://doi.org/10.1007/s11245-013-9171-1

Beissner, F., Meissner, K., Bar, K. J., & Napadow, V. (2013). The autonomic brain: an activation likelihood estimation meta-analysis for central processing of autonomic function. *J NEUROSCI*, *33*(25), 10503-10511. https://doi.org/10.1523/JNEUROSCI.1103-13.2013

Beitz, C. (2001). Human rights as a common concern. *American Political Science Review*, 269-282.

Beitz, C. (2011). *The idea of human rights*. Oxford University Press.

Beitz, C. (2015). The force of subsistence rights. *Philosophical foundations of human rights*, 535-551.

Bell, J., & Strang, J. (2020). Medication Treatment of Opioid Use Disorder. *Biol Psychiatry*, *87*(1), 82-88. https://doi.org/10.1016/j.biopsych.2019.06.020

Bennabi, D., Pedron, S., Haffen, E., Monnin, J., Peterschmitt, Y., & Van Waes, V. (2014). Transcranial direct current stimulation for memory enhancement: from clinical research to animal models. *Front Syst Neurosci*, *8*, 159. https://doi.org/10.3389/fnsys.2014.00159

Bennett, C. (2008). The apology ritual: A philosophical theory of punishment.

Bennett, C. (2018). Intrusive intervention and opacity respect. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice* (pp. 255-273). Oxford University Press.

Berger, T. W., Hampson, R. E., Song, D., Goonawardena, A., Marmarelis, V. Z., & Deadwyler, S. A. (2011). A cortical neural prosthesis for restoring and enhancing memory. *J Neural Eng*, *8*(4), 046017. https://doi.org/10.1088/1741-2560/8/4/046017

Berlin, I. (1969). *Four essays on liberty*. Oxford University Press.

Berman, G., & Dar, A. (2013). Prison population statistics. *London: House of Commons Library*.

Berman, M. E., McCloskey, M. S., Fanning, J. R., Schumacher, J. A., & Coccaro, E. F. (2009). Serotonin augmentation reduces response to attack in aggressive individuals. *Psychol Sci*, *20*(6), 714-720. https://doi.org/10.1111/j.1467-9280.2009.02355.x

Berns, G. S., Laibson, D., & Loewenstein, G. (2007). Intertemporal choice--toward an integrative framework. *TRENDS COGN SCI*, *11*(11), 482-488. https://doi.org/10.1016/j.tics.2007.08.011

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Rao, S. M., & Cox, R. W. (1999). Conceptual processing during the conscious resting state. A functional MRI study. *J Cogn Neurosci*, *11*(1), 80-95. https://doi.org/10.1162/089892999563265

Birks, D. (2018). Can Neurointerventions Communicate Censure? In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press.

Birks, D., & Buyx, A. (2018). Punishing Intentions and Neurointerventions. *AJOB neuroscience*, *9*(3), 133-143. https://doi.org/10.1080/21507740.2018.1496162

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

Birks, D., & Douglas, T. (2018). *Treatment for crime : philosophical essays on neurointerventions in criminal justice*. Oxford : Oxford University Press; First edition.

Blair, J., Sellars, C., Strickland, I., Clark, F., Williams, A., Smith, M., & Jones, L. (1996). Theory of mind in the psychopath. *Journal of Forensic Psychiatry*, *7*(1), 15-25.

Blair, R. J. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *TRENDS COGN SCI*, *11*(9), 387-392. https://doi.org/10.1016/j.tics.2007.07.003

Blitz, M. J. (2010). Freedom of thought for the extended mind: Cognitive enhancement and the constitution. *WISC LAW REV*, *2010*(4), 1049-1117.

Bokulich, A. (2001). Rethinking thought experiments. *Perspectives on Science*, *9*(3), 285-307.

Bomann-Larsen, L. (2013). Voluntary Rehabilitation? On Neurotechnological Behavioural Treatment, Valid Consent and (In)appropriate Offers. *Neuroethics*, *6*(1), 65-77. https://doi.org/10.1007/s12152-011-9105-9

Bongiovi, J. R. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power by Shoshana Zuboff. *Social Forces*, *98*(2), 23-23.

BonJour, L. (1998). *In defense of pure reason: A rationalist account of a priori justification*. Cambridge University Press.

Boonin, D. (2008). *The problem of punishment*. Cambridge University Press Cambridge.

Bosoer, L. (2021). Opinion: Chile at the Forefront of Neurorights Protection. https://blogs.eui.eu/latin-american-working-group/opinion-chile-at-the-forefront-of-neurorights-protection/

Bostrom, N., & Ord, T. (2006). The reversal test: eliminating status quo bias in applied ethics. *Ethics*, *116*(4), 656-679. https://doi.org/10.1086/505233

Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., & Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *NAT NEUROSCI*, *8*(9), 1263-1268. https://doi.org/10.1038/nn1525

Bradley, A. B. (2018). *Ending overcriminalization and mass incarceration : hope from civil society*. Cambridge : Cambridge University Press.

Brecker, K., Lins, S., & Sunyaev, A. (2023). Why it Remains Challenging to Assess Artificial Intelligence. 56th Hawaii Conference on System Sciences (HICSS),

Bremner, J. D. (2022). Traumatic stress: effects on the brain. *Dialogues in clinical neuroscience*.

Brent, M. (2020). Agent causation as a solution to the problem of action. *Canadian Journal of Philosophy*, *47*(5), 656-673. https://doi.org/10.1080/00455091.2017.1285643

Brevet-Aeby, C., Brunelin, J., Iceta, S., Padovan, C., & Poulet, E. (2016). Prefrontal cortex and impulsivity: Interest of noninvasive brain stimulation. *Neurosci Biobehav Rev*, *71*, 112-134. https://doi.org/10.1016/j.neubiorev.2016.08.028

Brink, D. O. (2010). *Moral realism and the foundations of ethics*. Cambridge University Press.

Bronstein, J. (2010). Objecting to the Genetic Virtue Program. *Politics Life Sci*, *29*(1), 85-87. https://doi.org/10.2990/29_1_85

Brooks, T. (2021). *Punishment* (2nd ed.). Routledge. https://doi.org/https://doi.org/10.4324/9780203929421

Brown, D. K. (2002). What virtue ethics can do for criminal justice: A reply to Huigens. *Wake Forest L. Rev.*, *37*, 29.

Brown, T. (2015). A Relational Take on Advisory Brain Implant Systems. *AJOB neuroscience*, *6*(4), 46-47. https://doi.org/10.1080/21507740.2015.1094559

Bruhl, A. B., d'Angelo, C., & Sahakian, B. J. (2019). Neuroethical issues in cognitive enhancement: Modafinil as the example of a workplace drug? *Brain Neurosci Adv*, *3*, 2398212818816018. https://doi.org/10.1177/2398212818816018

Bruno, M.-A., Fernández-Espejo, D., Lehembre, R., Tshibanda, L., Vanhaudenhuyse, A., Gosseries, O., Lommers, E., Napolitani, M., Noirhomme, Q., & Boly, M. (2011). Multimodal neuroimaging in patients with disorders of consciousness showing "functional hemispherectomy". *Progress in brain research*, *193*, 323-333.

Bublitz, J.-C. (2013). My mind is mine!? Cognitive liberty as a legal concept. In *Cognitive enhancement* (pp. 233-264). Springer Netherlands.

Bublitz, J.-C. (2014). Freedom of thought in the age of neuroscience. *Archiv für Rechts-und Sozialphilosphie*, *100*(1), 1-25.

Bublitz, J.-C. (2015a). Cognitive Liberty or the International Human Right to Freedom of Thought. In J. Clausen & N. Levy (Eds.), *Handbook of Neuroethics* (pp. 1309-1333). Springer Netherlands. https://doi.org/10.1007/978-94-007-4707-4_166

Bublitz, J.-C. (2015b). Moral Enhancement and Mental Freedom. *Journal of applied philosophy*, *33*(1), 88-106. https://doi.org/10.1111/japp.12108

Bublitz, J.-C. (2018). "The Soul is the Prison of the Body"—Mandatory Moral Enhancement, Punishment and Rights Against Neuro-Rehabilitation. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice* (pp. 289-320). Oxford University Press.

Bublitz, J.-C. (2020a). The Nascent Right to Psychological Integrity and Mental Self-Determination. In A. Von Arnauld, Von der Decken, & K. M. Susi (Eds.), *The Cambridge Handbook of New Human Rights* (pp. 387-403). Cambridge University Press. https://doi.org/10.1017/9781108676106.031

Bublitz, J.-C. (2020b). Why Means Matter - Legally Relevant Differences Between Direct and Indirect Interventions into Other Minds. In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity* (pp. 49-159). Oxford University Press.

Bublitz, J.-C., & Merkel, R. (2014). Crimes against minds: on mental manipulations, harms and a human right to mental self-determination. *Criminal Law and Philosophy*, *8*(1), 51-77.

Buchanan, A. (2007a). Institutions, beliefs and ethics: Eugenics as a case study. *Journal of political philosophy*, *15*(1), 22-45.

Buchanan, A. (2007b). *Justice, legitimacy, and self-determination: Moral foundations for international law*. Oxford University Press on Demand.

Buchanan, A. (2009). Human nature and enhancement. *BIOETHICS*, *23*(3), 141-150. https://doi.org/10.1111/j.1467-8519.2008.00633.x

Buchanan, A. (2011). *Beyond humanity? : the ethics of biomedical enhancement*. Oxford University Press.

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655-661.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Ann N Y Acad Sci*, *1124*(1), 1-38. https://doi.org/10.1196/annals.1440.011

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *TRENDS COGN SCI*, *11*(2), 49-57. https://doi.org/10.1016/j.tics.2006.11.004

Bullock, E. (2018). Moral Paternalism and Neurointerventions. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press.

Bülow, W. (2020). "It Will Help You Repent". In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity*. Oxford University Press.

Burgess, A. (1962). *A Clockwork Orange*. William Heinemann.

Burke, J. F., Merkow, M. B., Jacobs, J., Kahana, M. J., & Zaghloul, K. A. (2014). Brain computer interface to enhance episodic memory in human participants. *FRONT HUM NEUROSCI*, *8*, 1055. https://doi.org/10.3389/fnhum.2014.01055

Butson, C. R., & McIntyre, C. C. (2006). Role of electrode design on the volume of tissue activated during deep brain stimulation. *J Neural Eng*, *3*(1), 1-8. https://doi.org/10.1088/1741-2560/3/1/001

Caldas, C. P., & Bertero, C. (2012). A concept analysis about temporality and its applicability in nursing care. *Nurs Forum*, *47*(4), 245-252. https://doi.org/10.1111/j.1744-6198.2012.00277.x

Campbell, J. (2008). *The hero with a thousand faces*. Novato, Calif. : New World Library; 3rd ed.

Campbell, J. M., Huang, Z., Zhang, J., Wu, X., Qin, P., Northoff, G., Mashour, G. A., & Hudetz, A. G. (2020). Pharmacologically informed machine learning approach for identifying pathological states of unconsciousness via resting-state fMRI. *neuroimage*, *206*, 116316. https://doi.org/10.1016/j.neuroimage.2019.116316

Campbell, K. (2017). Legal Rights. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/legal-rights/

Canada, J. H. S. o. (2018). Financial facts on Canadian prisons. https://johnhoward.ca/blog/financial-facts-canadian-prisons/

Caney, S. (2009). Climate change and the future: Discounting for time, wealth, and risk. *Journal of social philosophy*, *40*(2), 163-186.

Canli, T. (2015). Neurogenethics: An emerging discipline at the intersection of ethics, neuroscience, and genomics. *Appl Transl Genom*, *5*, 18-22. https://doi.org/10.1016/j.atg.2015.05.002

Canton, R. (2017). *Why Punish?: An Introduction to the Philosophy of Punishment*. Macmillan International Higher Education.

Caplan, A. L. (2006). Ethical issues surrounding forced, mandated, or coerced treatment. *J Subst Abuse Treat*, *31*(2), 117-120. https://doi.org/10.1016/j.jsat.2006.06.009

Caplan, A. L. (2008). Editorial. *Addiction*, *103*, 1919-1921.

Carlsmith, K. M. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research*, *21*(2), 119-137.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of personality and social psychology*, *83*(2), 284.

Carlson, E. A. (2001). *The unfit : a history of a bad idea*. Cold Spring Harbor, N.Y. : Cold Spring Harbor Laboratory Press.

Carman, M. (2021). The limits of direct modulation of emotion for moral enhancement. *BIOETHICS*, *35*(2), 192-198. https://doi.org/10.1111/bioe.12800

Carnahan, T., & McFarland, S. (2007). Revisiting the Stanford prison experiment: could participant self-selection have led to the cruelty? *Pers Soc Psychol Bull*, *33*(5), 603-614. https://doi.org/10.1177/0146167206292689

Caviola, L., & Faber, N. S. (2015). Pills or Push-Ups? Effectiveness and Public Perception of Pharmacological and Non-Pharmacological Cognitive Enhancement. *FRONT PSYCHOL*, *6*, 1852. https://doi.org/10.3389/fpsyg.2015.01852

Caviola, L., Mannino, A., Savulescu, J., & Faulmuller, N. (2014). Cognitive biases can affect moral intuitions about cognitive enhancement. *Front Syst Neurosci*, *8*, 195. https://doi.org/10.3389/fnsys.2014.00195

Čehajić, S., Brown, R., & González, R. (2009). What do I care? Perceived ingroup responsibility and dehumanization as predictors of empathy felt for the victim group. *Group Processes & Intergroup Relations*, *12*(6), 715-729.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness studies*, *2*(3), 200-219.

Chalmers, D. (1996a). *The conscious mind : in search of a fundamental theory*. Oxford University Press.

Chalmers, D. (1996b). *The Conscious Mind: In Search of a Fundamental Theory*. Cary: Oxford University Press USA - OSO.

Chalmers, D. (2007). The hard problem of consciousness. *The Blackwell companion to consciousness*, 225-235.

Chandler, J. A. (2018). Neurolaw and Neuroethics. *Camb Q Healthc Ethics*, *27*(4), 590-598. https://doi.org/10.1017/S0963180118000117

Chandler, R. K., Fletcher, B. W., & Volkow, N. D. (2009). Treating drug abuse and addiction in the criminal justice system: improving public health and safety. *Jama*, *301*(2), 183-190. https://doi.org/10.1001/jama.2008.976

Changeux, J. P., Courrege, P., & Danchin, A. (1973). A theory of the epigenesis of neuronal networks by selective stabilization of synapses. *Proc Natl Acad Sci U S A*, *70*(10), 2974-2978. https://doi.org/10.1073/pnas.70.10.2974

Chekroud, A. M., Everett, J. A., Bridge, H., & Hewstone, M. (2014). A review of neuroimaging studies of race-related prejudice: does amygdala response reflect threat? *FRONT HUM NEUROSCI*, *8*, 179. https://doi.org/10.3389/fnhum.2014.00179

Chen, C. Y., Raine, A., Chou, K. H., Chen, I. Y., Hung, D., & Lin, C. P. (2016). Abnormal white matter integrity in rapists as indicated by diffusion tensor imaging. *BMC Neurosci*, *17*(1), 45. https://doi.org/10.1186/s12868-016-0278-3

Chew, C., Douglas, T., & Faber, N. S. (2018). Biological Interventions for Crime Prevention. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press. https://doi.org/DOI: 10.1093/oso/9780198758617.003.0002

Chhatbar, P. Y., & Feng, W. (2015). Data Synthesis in Meta-Analysis may Conclude Differently on Cognitive Effect From Transcranial Direct Current Stimulation. *Brain Stimul*, *8*(5), 974-976. https://doi.org/10.1016/j.brs.2015.06.001

Chiesa, A., Calati, R., & Serretti, A. (2011). Does mindfulness training improve cognitive abilities? A systematic review of neuropsychological findings. *CLIN PSYCHOL REV*, *31*(3), 449-464. https://doi.org/10.1016/j.cpr.2010.11.003

Chiong, W. (2020). Insiders and Outsiders: Lessons for Neuroethics from the History of Bioethics. *AJOB Neurosci*, *11*(3), 155-166. https://doi.org/10.1080/21507740.2020.1778118

Choudhury, S., & Slaby, J. (2016). *Critical neuroscience: A handbook of the social and cultural contexts of neuroscience*. John Wiley & Sons.

Choy, O., Focquaert, F., & Raine, A. (2018). Benign Biological Interventions to Reduce Offending. *Neuroethics*, *13*(1), 29-41. https://doi.org/10.1007/s12152-018-9360-0

Choy, O., Raine, A., & Hamilton, R. H. (2018). Stimulation of the Prefrontal Cortex Reduces Intentions to Commit Aggression: A Randomized, Double-Blind, Placebo-Controlled, Stratified, Parallel-Group Trial. *J NEUROSCI*, *38*(29), 6505-6512. https://doi.org/10.1523/JNEUROSCI.3317-17.2018

Christen, M., & Müller, S. (2017). The Ethics of Expanding Applications of Deep Brain Stimulation. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics* (pp. 51-65). Routledge. https://doi.org/10.4324/9781315708652-6

Christman, J. (2004). Relational Autonomy, Liberal Individualism, and the Social Constitution of Selves. *Philosophical Studies*, *117*(1/2), 143-164. https://doi.org/10.1023/B:PHIL.0000014532.56866.5c

Christman, J. (2009). *The Politics of Persons*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511635571

Churchland, P. S. (1989). *Neurophilosophy: Toward a unified science of the mind-brain*. MIT press.

Churchland, P. S. (2011). *Braintrust*. Princeton University Press.

Ciaramelli, E., Muccioli, M., Ladavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Soc Cogn Affect Neurosci*, *2*(2), 84-92. https://doi.org/10.1093/scan/nsm001

Citherlet, T., Crettaz von Roten, F., Kayser, B., & Guex, K. (2021). Acute Effects of the Wim Hof Breathing Method on Repeated Sprint Ability: A Pilot Study. *Front Sports Act Living*, *3*, 700757. https://doi.org/10.3389/fspor.2021.700757

Clark, A. (2008a). *Supersizing the mind : embodiment, action, and cognitive extension*. Oxford University Press.

Clark, A. (2008b). *Supersizing the mind: Embodiment, action, and cognitive extension*. OUP USA.

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7-19. https://doi.org/10.1093/analys/58.1.7

Clark, D. B., Cornelius, J., Wood, D. S., & Vanyukov, M. (2004). Psychopathology risk transmission in children of parents with substance use disorders. *Am J Psychiatry*, *161*(4), 685-691. https://doi.org/10.1176/appi.ajp.161.4.685

Clarke, R. K. (2003). *Libertarian accounts of free will*. Oxford University Press.

Clausen, J., & Levy, N. (2015). What is Neuroethics? In C. J. & L. N. (Eds.), *Handbook of Neuroethics* (pp. v–vii). Springer.

Coffman, B. A., Clark, V. P., & Parasuraman, R. (2014). Battery powered thought: enhancement of attention, learning, and memory in healthy adults using transcranial direct current stimulation. *neuroimage*, *85 Pt 3*, 895-908. https://doi.org/10.1016/j.neuroimage.2013.07.083

Comfort, N. (2009). The prisoner as model organism: malaria research at Stateville Penitentiary. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *40*(3), 190-203.

Conan, G. M. (2020). Frequently overlooked realistic moral bioenhancement interventions. *J Med Ethics*, *46*(1), 43-47. https://doi.org/10.1136/medethics-2019-105534

Conrad, P., & Leiter, V. (1981). *Sociology of Health and Illness: Critical Perspectives (First through to Ninth Editions)*. Worth Publishers.

Conti, C. L., & Nakamura-Palacios, E. M. (2014). Bilateral transcranial direct current stimulation over dorsolateral prefrontal cortex changes the drug-cued reactivity in the anterior cingulate cortex of crack-cocaine addicts. *Brain Stimul*, *7*(1), 130-132. https://doi.org/10.1016/j.brs.2013.09.007

Cooke, D. J., & Michie, C. (2001). Refining the construct of psychopathy: towards a hierarchical model. *Psychological assessment*, *13*(2), 171.

Cooter, R. (2014). Neural veils and the will to historical critique: why historians of science need to take the neuro-turn seriously. *Isis*, *105*(1), 145-154. https://doi.org/10.1086/675556

Coppola, F. (2018). Valuing Emotions in Punishment: an Argument for Social Rehabilitation with the Aid of Social and Affective Neuroscience. *Neuroethics*, *14*(S3), 251-268. https://doi.org/10.1007/s12152-018-9393-4

Cortese, S., Kelly, C., Chabernaud, C., Proal, E., Di Martino, A., Milham, M. P., & Castellanos, F. X. (2012). Toward systems neuroscience of ADHD: a meta-analysis of 55 fMRI studies. *Am J Psychiatry*, *169*(10), 1038-1055. https://doi.org/10.1176/appi.ajp.2012.11101521

Coward, L. A. (2013). *Towards a Theoretical Neuroscience: from Cell Chemistry to Cognition*. Dordrecht : Springer Netherlands : Imprint: Springer; 1st ed. 2013.

Cozolino, L. (2014). *The neuroscience of human relationships: Attachment and the developing social brain (Norton series on interpersonal neurobiology)*. WW Norton & Company.

Craig, J. N. (2016). Incarceration, Direct Brain Intervention, and the Right to Mental Integrity – a Reply to Thomas Douglas. *Neuroethics*, *9*(2), 107-118. https://doi.org/10.1007/s12152-016-9255-x

Crane, T. (2015). *The mechanical mind: A philosophical introduction to minds, machines and mental representation*. Routledge.

Critchley, H. D., & Harrison, N. A. (2013). Visceral influences on brain and behavior. *Neuron*, *77*(4), 624-638. https://doi.org/10.1016/j.neuron.2013.02.008

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proc Natl Acad Sci U S A*, *107*(40), 17433-17438. https://doi.org/10.1073/pnas.1009396107

Crockett, M. J., & Rini, R. A. (2015). Neuromodulators and the instability of moral cognition. In J. Decety & T. Wheatley (Eds.), *The Moral Brain: A Multidisciplinary Perspective* (pp. 221–235). MIT Press.

Cullen, F. T., Jonson, C. L., & Nagin, D. S. (2011). Prisons Do Not Reduce Recidivism. *The Prison Journal*, *91*(3_suppl), 48S-65S. https://doi.org/10.1177/0032885511415224

Currie, E. (1989). *Confronting crime: an American challenge*. Pantheon.

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353-380. https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F. (2013). Action, outcome, and value: a dual-system framework for morality. *Pers Soc Psychol Rev*, *17*(3), 273-292. https://doi.org/10.1177/1088868313495594

Cushman, F. (2015). Punishment in humans: From intuitions to institutions. *Philosophy Compass*, *10*(2), 117-133.

D'Angiulli, A., Herdman, A., Stapells, D., & Hertzman, C. (2008). Children's event-related potentials of auditory selective attention vary with their socioeconomic status. *Neuropsychology*, *22*(3), 293.

Dahl, C. J., Lutz, A., & Davidson, R. J. (2015). Reconstructing and deconstructing the self: cognitive mechanisms in meditation practice. *TRENDS COGN SCI*, *19*(9), 515-523. https://doi.org/10.1016/j.tics.2015.07.001

Damasio, A. R. (1994). *Descartes' error : emotion, reason, and the human brain*. New York : Avon Books.

Damasio, A. R., & Damasio, H. (2021). *Feeling & knowing : making minds conscious* (First edition. ed.). Pantheon Books,.

Dambacher, F., Schuhmann, T., Lobbestael, J., Arntz, A., Brugman, S., & Sack, A. T. (2015). Reducing proactive aggression through non-invasive brain stimulation. *Soc Cogn Affect Neurosci*, *10*(10), 1303-1309. https://doi.org/10.1093/scan/nsv018

Dancy, J. (1993). *Moral reasons*. Blackwell Publishers.

Dancy, J. (2023). The Roots of Normativity, by Joseph Raz. *Mind*. https://doi.org/10.1093/mind/fzac069

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proc Natl Acad Sci U S A*, *108*(17), 6889-6892. https://doi.org/10.1073/pnas.1018033108

Darabont, F., & King, S. (1994a). The Shawshank Redemption. In. United States: Columbia TriStar Home Video.

Darabont, F., & King, S. (1994b). *The Shawshank Redemption* Columbia TriStar Home Video.

Darabont, F., & King, S. (1996). *The Shawshank Redemption: The Shooting Script*. Nick Hern Books. http://books.google.ca/books?id=CJqYPwAACAAJ

Darley, J. M., Carlsmith, K. M., & Robinson, P. H. (2000). Incapacitation and just deserts as motives for punishment. *Law and Human Behavior*, *24*(6), 659-683.

Darmani, G., Bergmann, T. O., Butts Pauly, K., Caskey, C. F., de Lecea, L., Fomenko, A., Fouragnan, E., Legon, W., Murphy, K. R., Nandi, T., Phipps, M. A., Pinton, G., Ramezanpour, H., Sallet, J., Yaakub, S. N., Yoo, S. S., & Chen, R. (2022). Non-invasive transcranial ultrasound stimulation for neuromodulation. *Clinical Neurophysiology*, *135*, 51-73. https://doi.org/10.1016/j.clinph.2021.12.010

Darwall, S. (1997). *Moral discourse and practice : some philosophical approaches*. New York : Oxford University Press.

Darwall, S. (2006). The Value of Autonomy and Autonomy of the Will. *Ethics*, *116*(2), 263-284. https://doi.org/10.1086/498461

Darwall, S. L. (1977). Two Kinds of Respect. *Ethics*, *88*(1), 36-49. https://doi.org/10.1086/292054

Darwin, C. (1860/2010). *The origin of species: A variorum text*. University of Pennsylvania Press.

Davidson, D. (2001a). *Essays on actions and events* (2nd ed.). Clarendon.

Davidson, D. (2001b). How Is Weakness of the Will Possible? In. Oxford University Press. https://doi.org/10.1093/0199246270.003.0002

Davidson, R. J., Jackson, D. C., & Kalin, N. H. (2000). Emotion, plasticity, context, and regulation: perspectives from affective neuroscience. *PSYCHOL BULL*, *126*(6), 890-909. https://doi.org/10.1037/0033-2909.126.6.890

Davidson, R. J., & McEwen, B. S. (2012). Social influences on neuroplasticity: stress and interventions to promote well-being. *NAT NEUROSCI*, *15*(5), 689-695. https://doi.org/10.1038/nn.3093

Davies, P. S. (2009). *Subjects of the World: Darwin's Rhetoric and the Study of Agency in Nature*. The University of Chicago Press.

Davies, P. S. (2020). Foundational Facts for Legal Responsibility - Human Agency and the Aims of Restorative Neurointerventions. In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity* (pp. 319-348). Oxford University Press.

Davis, F. C., Knodt, A. R., Sporns, O., Lahey, B. B., Zald, D. H., Brigidi, B. D., & Hariri, A. R. (2013). Impulsivity and the modular organization of resting-state neural networks. *CEREB CORTEX*, *23*(6), 1444-1452. https://doi.org/10.1093/cercor/bhs126

Davis, M. (1992). *To make the punishment fit the crime: essays in the theory of criminal justice*. Westview Pr.

De Deyn, P. P., & Buitelaar, J. (2006). Risperidone in the management of agitation and aggression associated with psychiatric disorders. *Eur Psychiatry*, *21*(1), 21-28. https://doi.org/10.1016/j.eurpsy.2005.11.003

De Gelder, B., Snyder, J., Greve, D., Gerard, G., & Hadjikhani, N. (2004). Fear fosters flight: a mechanism for fear contagion when perceiving emotion expressed by a whole body. *Proceedings of the National Academy of Sciences*, *101*(47), 16701-16706.

de Haan, S., Rietveld, E., Stokhof, M., & Denys, D. (2015). Effects of Deep Brain Stimulation on the Lived Experience of Obsessive-Compulsive Disorder Patients:

In-Depth Interviews with 18 Patients. *PLOS ONE*, *10*(8), e0135524. https://doi.org/10.1371/journal.pone.0135524

de Melo-Martin, I., & Salles, A. (2015). Moral bioenhancement: much ado about nothing? *BIOETHICS*, *29*(4), 223-232. https://doi.org/10.1111/bioe.12100

Decety, J., & Howard, L. H. (2013). The role of affect in the neurodevelopment of morality. *Child Development Perspectives*, *7*(1), 49-54.

Decety, J., Skelly, L. R., & Kiehl, K. A. (2013). Brain response to empathy-eliciting scenarios involving pain in incarcerated individuals with psychopathy. *JAMA Psychiatry*, *70*(6), 638-645. https://doi.org/10.1001/jamapsychiatry.2013.27

Deckers, L. (2014). *Motivation : biological, psychological, and environmental*. Boston : Pearson; Fourth edition.

Dedoncker, J., Brunoni, A. R., Baeken, C., & Vanderhasselt, M. A. (2016). A Systematic Review and Meta-Analysis of the Effects of Transcranial Direct Current Stimulation (tDCS) Over the Dorsolateral Prefrontal Cortex in Healthy and Neuropsychiatric Samples: Influence of Stimulation Parameters. *Brain Stimul*, *9*(4), 501-517. https://doi.org/10.1016/j.brs.2016.04.006

DeGrazia, D. (2005). *Human identity and bioethics*. Cambridge University Press. Table of contents only http://www.loc.gov/catdir/toc/ecip051/2004021782.html

Publisher description http://www.loc.gov/catdir/enhancements/fy0632/2004021782-d.html

DeGrazia, D. (2014). Moral enhancement, freedom, and what we (should) value in moral behaviour. *Journal of medical ethics*, *40*(6), 361-368.

Dehaene, S. (2014). *Consciousness and the brain : deciphering how the brain codes our thoughts*. New York, New York : Viking.

Dehaene, S., & Changeux, J.-P. (2004). Neural mechanisms for access to consciousnes. In M. Gazzanig (Ed.), *The cognitive neurosciences* MIT Press.

Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200-227. https://doi.org/10.1016/j.neuron.2011.03.018

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1-37. https://doi.org/10.1016/s0010-0277(00)00123-2

Deisseroth, K. (2011). Optogenetics. *NAT METHODS*, *8*(1), 26-29. https://doi.org/10.1038/nmeth.f.324

Delgado-Herrera, M., Reyes-Aguilar, A., & Giordano, M. (2021). What Deception Tasks Used in the Lab Really Do: Systematic Review and Meta-analysis of Ecological Validity of fMRI Deception Tasks. *Neuroscience*, *468*, 88-109. https://doi.org/10.1016/j.neuroscience.2021.06.005

Deming, P., & Koenigs, M. (2020). Functional neural correlates of psychopathy: a meta-analysis of MRI data. *Transl Psychiatry*, *10*(1), 133. https://doi.org/10.1038/s41398-020-0816-8

Dennett, D. (1990). Quining qualia. InW. Lycan (Ed.), Mind and cognition: A reader (pp. 519 547). In: Oxford: Blackwell.

Dennett, D. C. (1984). *Elbow room : the varieties of free will worth wanting*. Cambridge, Mass. : MIT Press.

Dennett, D. C. (1993). *Consciousness explained*. Penguin Books Limited.

Dennett, D. C. (2015). *Elbow room : the varieties of free will worth wanting*. Cambridge, Massachusetts ; London, England : MIT Press; New edition.

Denno, D. (2016). The Place for Neuroscience in Criminal Law. In D. Patterson & M. Pardo (Eds.), *Philosophical Foundations of Law and Neuroscience*. Oxford University Press.

Deonna, J., & Teroni, F. (2012). *The Emotions: A Philosophical Introduction*. Taylor and Francis. https://doi.org/10.4324/9780203721742

Derbyshire, S., & Raja, A. (2011). On the development of painful experience. *Journal of Consciousness studies*, *18*(9-10), 233-256.

Descartes, R. (1641/1996). *Discourse on the method: And, meditations on first philosophy*. Yale University Press.

DeVeaux, M. i. (2013). The trauma of the incarceration experience. *HARVARD CIVIL RIGHTS*, *48*(1), 257-277.

Devlin, P. (2009). *The enforcement of morals*. Indianapolis : Liberty Fund.

Dillon, R. (2018). Banning Solitary for Prisoners with Mental Illness: The Blurred Line Between Physical and Psychological Harm. *Nw. JL & Soc. Pol'y*, *14*, 265.

Dobbs, D. (2005). Fact or phrenology? *Scientific American Mind*, *16*(1), 24-31.

Dolan, K. A., Shearer, J., White, B., Zhou, J., Kaldor, J., & Wodak, A. D. (2005). Four-year follow-up of imprisoned male heroin users and methadone treatment: mortality, re-incarceration and hepatitis C infection. *Addiction*, *100*(6), 820-828. https://doi.org/10.1111/j.1360-0443.2005.01050.x

Dolan, P., Hallsworth, M., Halpern, D., King, D., & Vlaev, I. (2010). MINDSPACE: influencing behaviour for public policy.

Dolinko, D. (1991). Some thoughts about retributivism. *Ethics*, *101*(3), 537-559.

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press.

Donnelly-Lazarov, B. (2021). Neuroscience and the Moral Enhancement of Offenders: The Exceptionally 'Good' Brain as a Thought Experiment. In *Neurolaw* (pp. 229-250). Springer International Publishing. https://doi.org/10.1007/978-3-030-69277-3_10

Doob, A. N., & Webster, C. M. (2003). Sentence severity and crime: Accepting the null hypothesis. *Crime and Justice*, *30*, 143-195.

Dostoyevsky, F. (1995 [1869]). *The idiot*. New York : Dramatists Play Service.

Dostoyevsky, F. ([1880] 2003). *The brothers karamazov*. Penguin UK.

Douglas, T. (2013). Moral enhancement via direct emotion modulation: a reply to John Harris. *BIOETHICS*, *27*(3), 160-168. https://doi.org/10.1111/j.1467-8519.2011.01919.x

Douglas, T. (2014a). Blurred lines: Neurointerventions in crime prevention. *The University of Otago Magazin*. http://www.otago.ac.nz/bioethics/news/otago082730.html

Douglas, T. (2014b). Blurred lines: Neurointerventions in crime prevention. *The University of Otago Magazin*. http://www.otago.ac.nz/bioethics/news/otago082730.html

Douglas, T. (2014c). Criminal rehabilitation through medical intervention: moral liability and the right to bodily integrity. *The Journal of Ethics*, *18*(2), 101-122.

Douglas, T. (2014d). Enhancing Moral Conformity and Enhancing Moral Worth. *Neuroethics*, *7*(1), 75-91. https://doi.org/10.1007/s12152-013-9183-y

Douglas, T. (2018). Neural and Environmental Modulation of Motivation. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice* (pp. 208-223). Oxford University Press.

Douglas, T., Bonte, P., Focquaert, F., Devolder, K., & Sterckx, S. (2013). Coercion, incarceration, and chemical castration: an argument from autonomy. *J Bioeth Inq*, *10*(3), 393-405. https://doi.org/10.1007/s11673-013-9465-4

Dowling, J. E. (1998). *Creating mind: How the brain works*. WW Norton & Company.

Drane, D., Swarat, S., Light, G., Hersam, M., & Mason, T. (2009). An evaluation of the efficacy and transferability of a nanoscience module. *Journal of Nano Education*, *1*(1), 8-14.

Draper, I. T. (1974). The Working Brain (an Introduction to Neuropsychology). *Journal of Neurology, Neurosurgery & Psychiatry*, *37*(3), 361-362. https://doi.org/10.1136/jnnp.37.3.361-b

Dresler, M., Sandberg, A., Bublitz, J.-C., Ohla, K., Trenado, C., Mroczko-Wasowicz, A., Kuhn, S., & Repantis, D. (2019). Hacking the Brain: Dimensions of Cognitive Enhancement. *ACS Chem Neurosci*, *10*(3), 1137-1148. https://doi.org/10.1021/acschemneuro.8b00571

Dresp-Langley, B. (2020). Seven properties of self-organization in the human brain. *Big Data and Cognitive Computing*, *4*(2), 10.

Dubljević, V. (2013). Autonomy in neuroethics: Political and not metaphysical. *AJOB neuroscience*, *4*(4), 44-51.

Dubljević, V. (2016). Autonomy is Political, Pragmatic, and Postmetaphysical: A Reply to Open Peer Commentaries on "Autonomy in Neuroethics". *AJOB neuroscience*, *7*(4), W1-W3.

Dubljevic, V., & Racine, E. (2014). A single cognitive heuristic process meets the complexity of domain-specific moral heuristics. *Behav Brain Sci*, *37*(5), 487-488. https://doi.org/10.1017/S0140525X13003701

Dubljevic, V., & Racine, E. (2017). Moral Enhancement Meets Normative and Empirical Reality: Assessing the Practical Feasibility of Moral Enhancement Neurotechnologies. *BIOETHICS*, *31*(5), 338-348. https://doi.org/10.1111/bioe.12355

Dubljevic, V., Saigle, V., & Racine, E. (2014). The rising tide of tDCS in the media and academic literature. *Neuron*, *82*(4), 731-736. https://doi.org/10.1016/j.neuron.2014.05.003

Duff, A. (1991). *Trials and punishments*. CUP Archive.

Duff, A. (2007). *Answering for crime: Responsibility and liability in the criminal law*. Bloomsbury Publishing.

Duff, A., & Duff, R. A. (2001). *Punishment, communication, and community*. Oxford University Press, USA.

Duff, A., & Hoskins, Z. (2019). Legal Punishment. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/legal-punishment/

Durose, M. R., Cooper, A. D., & Snyder, H. N. (2014). *Recidivism of prisoners released in 30 states in 2005: Patterns from 2005 to 2010* (Vol. 28). US Department of Justice, Office of Justice Programs, Bureau of Justice ….

Düzel, E., Guitart-Masip, M., Maass, A., Hämmerer, D., Betts, M. J., Speck, O., Weiskopf, N., & Kanowski, M. (2015). Midbrain fMRI: applications, limitations and challenges. *fMRI: From nuclear spins to brain functions*, 581-609.

Dworkin, G. (1988a). The Theory and Practice of Autonomy In. New York, NY: Cambridge University Press.

Dworkin, G. (1988b). *The Theory and Practice of Autonomy*. Cambridge University Press.

Dworkin, G. (2005). Moral paternalism. *Law and Philosophy*, *24*(3), 305-319.

Dworkin, R. (1977). Liberty and moralism. *Taking rights seriously*, 289-310.

Dworkin, R. (1985). Law's Ambitions for Itself. *Virginia law review*, *71*(2), 173-187. https://doi.org/10.2307/1073016

Dworkin, R. (1986). *Law's Empire*. Harvard University Press.

Dworkin, R. (1996). Objectivity and truth: You'd better believe it. *Philosophy & Public Affairs*, *25*(2), 87-139.

Dworkin, R. (2011). *Justice for Hedgehogs*. Harvard University Press.

Dworkin, R., & Waldron, J. (2013). Rights as trumps. *Arguing about law*, 335-344.

Eagleson, G., Waldersee, R., & Simmons, R. (2000). Leadership behaviour similarity as a basis of selection into a management team. *British Journal of Social Psychology*, *39*(2), 301-308.

Earp, B. D., Douglas, T., & Savulescu, J. (2017). Moral neuroenhancement. In S. Johnson & K. Rommelfanger (Eds.), *The Routledge handbook of neuroethics* (pp. 166-184). Routledge.

Earp, B. D., Douglas, T., & Savulescu, J. (2018). Moral Neuroenhancement. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Edmundson, W. A. (2012). *An Introduction to Rights*. Cambridge University Press. http://ebookcentral.proquest.com/lib/ucalgary-ebooks/detail.action?docID=862390

Egli, N., Pina, M., Christensen, P. S., Aebi, M., & Killias, M. (2009). Effects of drug substitution programs on offending among drug-addicts. *Campbell Systematic Reviews*, *5*(1), 1-40.

El-Hai, J. (2005). *The lobotomist: a maverick medical genius and his tragic quest to rid the world of mental illness*. J. Wiley Hoboken, NJ.

Eley, S., Gallop, K., McIvor, G., Morgan, K., & Yates, R. (2002). Drug treatment and testing orders: Evaluation of the Scottish pilots. *Scottish Executive Central Research Unit*.

Elfferich, A. M. (2021). Social Justice Theories as the Basis for Public Policy on Psychopharmacological Cognitive Enhancement [Article]. *Canadian Journal of Bioethics (Revue canadienne de bioéthique)*, *4*, 126-136. https://link-gale-com.ezproxy.lib.ucalgary.ca/apps/doc/A668712664/CPI?u=ucalgary&sid=bookmark-CPI&xid=f87fbf58

Elliott, C. (1999). Bioethics, culture and identity: A philosophical disease. In: New York: Routledge.

Elliott, C. (2003). *Better than well : American medicine meets the American dream*. New York : W.W. Norton; 1st ed.

Elster, J. (2015). *Explaining social behavior: More nuts and bolts for the social sciences*. Cambridge University Press.

Laidlaw, E. *Technology-Facilitated Mind Hacking: Protection of Inner Freedoms in Canadian Law*. Centre for International Governance Innovation.

Erler, A. (2020). Neuroenhancement, Coercion, and Neo-Luddism. In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity*. Oxford University Press.

Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A bad taste in the mouth: gustatory disgust influences moral judgment. *Psychol Sci*, *22*(3), 295-299. https://doi.org/10.1177/0956797611398497

Evans, J. S. B., & Frankish, K. E. (2009). *In two minds: Dual processes and beyond*. Oxford University Press.

Facco, E., Agrillo, C., & Greyson, B. (2015). Epistemological implications of near-death experiences and other non-ordinary mental expressions: Moving beyond the concept of altered state of consciousness. *Med Hypotheses*, *85*(1), 85-93. https://doi.org/10.1016/j.mehy.2015.04.004

Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent*. Oxford University Press.

Fallon, J. (2014). *The psychopath inside: A neuroscientist's personal journey into the dark side of the brain*. Penguin.

Fallon, S. J., van der Schaaf, M. E., Ter Huurne, N., & Cools, R. (2017). The Neurocognitive Cost of Enhancing Cognition with Methylphenidate: Improved Distractor Resistance but Impaired Updating. *J Cogn Neurosci*, *29*(4), 652-663. https://doi.org/10.1162/jocn_a_01065

Farah, M. J. (2011a). Neuroscience and Neuroethics in the 21st Century. In J. Illes & B. Sahakian (Eds.), *Oxford Handbook of Neuroethics* (pp. 761-781). Oxford University Press.

Farah, M. J. (2011b). Neuroscience and Neuroethics in the 21st Century. In J. Illes & B. Sahakian (Eds.), *Oxford Handbook of Neuroethics* (pp. 761-781). Oxford University Press.

Farah, M. J., Hutchinson, J. B., Phelps, E. A., & Wagner, A. D. (2014). Functional MRI-based lie detection: scientific and societal challenges. *NAT REV NEUROSCI*, *15*(2), 123-131. https://doi.org/10.1038/nrn3665

Farahany, N., & Ramos, K. M. (2020). Neuroethics: Fostering Collaborations to Enable Neuroscientific Discovery. *AJOB Neurosci*, *11*(3), 148-154. https://doi.org/10.1080/21507740.2020.1778117

Faria Jr, M. A. (2013). Violence, mental illness, and the brain–A brief history of psychosurgery: Part 1–From trephination to lobotomy. *Surgical Neurology International*, *4*, 49-49.

Fede, S. J., & Kiehl, K. A. (2020). Meta-analysis of the moral brain: patterns of neural engagement assessed using multilevel kernel density analysis. *Brain Imaging Behav*, *14*(2), 534-547. https://doi.org/10.1007/s11682-019-00035-5

Feigenson, N. (2007). Brain imaging and courtroom evidence: on the admissibility and persuasiveness of fMRI. *International journal of law in context*, *2*(3), 233-255. https://doi.org/10.1017/s174455230600303x

Feinberg, J. (1978). Pornography and the criminal law. *U. Pitt. L. Rev.*, *40*, 567.

Feinberg, J. (1986). Harm to Self (vol III): The Moral Limits of Criminal Law. In: New York: Oxford University Press.

Feinberg, J. (1987). *Harm to others* (Vol. 1). Oxford University Press on Demand.

Feinberg, J., & Narveson, J. (1970). The nature and value of rights. *The Journal of Value Inquiry*, *4*(4), 243-260.

Felten, D. L., O'Banion, M. K., & Maida, M. E. (2015). *Netter's atlas of neuroscience*. Elsevier Health Sciences.

Fincham, F. D., & Roberts, C. (1985). Intervening causation and the mitigation of responsibility for harm doing II. The role of limited mental capacities. *Journal of Experimental Social Psychology*, *21*(2), 178-194.

Fingarette, H. (2013). Punishment and suffering. *The American Philosophical Association Centennial Series*, 439-461.

Finke, K., Dodds, C. M., Bublak, P., Regenthal, R., Baumann, F., Manly, T., & Muller, U. (2010). Effects of modafinil and methylphenidate on visual attention capacity: a TVA-based study. *Psychopharmacology (Berl), 210*(3), 317-329. https://doi.org/10.1007/s00213-010-1823-x

Fischer, J. M. (1994). *The metaphysics of free will* (Vol. 1). Oxford: Blackwell.

Fischer, J. M. (1998). *Responsibility and control : a theory of moral responsibility*. Cambridge : Cambridge University Press.

Fischer, John M. (1999). Recent Work on Moral Responsibility. *Ethics*, *110*(1), 93-139. https://doi.org/10.1086/233206

Fischer, J. M. (2006a). The Cards that are Dealt You. *The Journal of Ethics*, *10*(1-2), 107-129. https://doi.org/10.1007/s10892-005-4594-6

Fischer, J. M. (2006b). *My way : essays on moral responsibility*. Oxford University Press.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control : a theory of moral responsibility*. Cambridge University Press.

Fischer, J. M., & Ravizza, M. (2000). *Responsibility and control: A theory of moral responsibility*. Cambridge university press.

Flathman, R. E. (1976). *The practice of rights*. Cambridge University Press.

Focquaert, F., Assche, K. V., & Sterckx, S. (2020). Offering Neurointerventions to Offenders With Cognitive-Emotional Impairments. In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity*. Oxford University Press.

Focquaert, F., & Schermer, M. (2015). Moral Enhancement: Do Means Matter Morally? *Neuroethics*, *8*(2), 139-151. https://doi.org/10.1007/s12152-015-9230-y

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford review*, *5*.

Foot, P. (2002). *Virtues and vices and other essays in moral philosophy*. Oxford University Press on Demand.

Forman, M. (1970). *One Flew Over the Cuckoo's Nest*. Samuel French Inc.

Forsberg, L. (2018). Crime-Preventing Neurointerventions and the Law. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press.

Foster, C. (2011). *Human dignity in bioethics and law*. Bloomsbury Publishing.

Foster, C. (2015). Human dignity in bioethics and law. *J Med Ethics*, *41*(12), 935. https://doi.org/10.1136/medethics-2013-101339

Foucault, M. ([1975] 2012). *Discipline and punish: The birth of the prison*. Vintage.

Fox, K. C., Nijeboer, S., Dixon, M. L., Floman, J. L., Ellamil, M., Rumak, S. P., Sedlmeier, P., & Christoff, K. (2014). Is meditation associated with altered brain structure? A systematic review and meta-analysis of morphometric neuroimaging in meditation practitioners. *Neurosci Biobehav Rev*, *43*, 48-73. https://doi.org/10.1016/j.neubiorev.2014.03.016

Fox, R. C., & Swazey, J. P. (2013). *Spare parts: Organ replacement in American society*. Transaction Publishers.

Fox, S. E., Levitt, P., & Nelson, C. A. (2010). How the timing and quality of early experiences influence the development of brain architecture. *Child development*, *81*(1), 28-40.

Franke, A. G., Bagusat, C., Rust, S., Engel, A., & Lieb, K. (2014). Substances used and prevalence rates of pharmacological cognitive enhancement among healthy subjects. *Eur Arch Psychiatry Clin Neurosci*, *264 Suppl 1*(S1), S83-90. https://doi.org/10.1007/s00406-014-0537-1

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of philosophy*, *66*(23), 829-839. https://doi.org/10.2307/2023833

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of philosophy*, *68*(1), 5-20. https://doi.org/10.2307/2024717

Frankfurt, H. G. (1988a). Coercion and moral responsibility. In H. Frankfurt (Ed.), *The importance of what we care about* (pp. 11-25). Cambridge University Press.

Frankfurt, H. G. (1988b). Free will and the concept of a person. In H. Frankfurt (Ed.), *The importance of what we care about* (pp. 11-25). Cambridge University Press.

Frankfurt, H. G. (1988c). *The Importance of What We Care About*. Cambridge University Press.

Frankfurt, H. G. (2018). Alternate possibilities and moral responsibility. In *Moral Responsibility and Alternative Possibilities* (pp. 17-25). Routledge.

Frankish, K. (2009). Systems and levels: Dual-system theories and the personalsubpersonal distinction. *J. St. BT Evans & K. Frankish*, 89-107.

Frankl, V. E. (1985). *Man's search for meaning*. Simon and Schuster.

Franzini, A., Broggi, G., Cordella, R., Dones, I., & Messina, G. (2013). Deep-brain stimulation for aggressive and disruptive behavior. *World Neurosurg*, *80*(3-4), S29 e11-24. https://doi.org/10.1016/j.wneu.2012.06.038

Frase, R. S. (2003). Limiting retributivism: The consensus model of criminal punishment. In *The future of imprisonment in the 21st century*. Oxford University Press.

Frederick, S., Loewenstein, G., & O'donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of economic literature*, *40*(2), 351-401.

Freeman, S. (2007). *Rawls*. Routledge.

Freud, S. (1899 [1965]). *The Interpretation of Dreams*. Avon Books.

Friedlander, H. (1995). *Origins of Nazi Genocide: From Euthanasia to the Final Solution*. Chapel Hill: The University of North Carolina Press.

Friedrich, O., Racine, E., Steinert, S., Pömsl, J., & Jox, R. J. (2018). An Analysis of the Impact of Brain-Computer Interfaces on Autonomy. *Neuroethics*, *14*(1), 17-29. https://doi.org/10.1007/s12152-018-9364-9

Friston, K. (2010). The free-energy principle: a unified brain theory? *NAT REV NEUROSCI*, *11*(2), 127-138. https://doi.org/10.1038/nrn2787

Fuchs, T. (2004). The challenge of neuroscience: Psychiatry and phenomenology today. In.

Fuchs, T. (2008). The brain as a relational organ: a phenomenological and ecological concept. *Kohlhammer, Stuttgart*.

Fuchs, T. (2011). The brain--A mediating organ. *Journal of Consciousness studies*, *18*(7-8), 196-221.

Fuchs, T., & Schlimme, J. E. (2009). Embodiment and psychopathology: a phenomenological perspective. *Curr Opin Psychiatry*, *22*(6), 570-575. https://doi.org/10.1097/YCO.0b013e3283318e5c

Fukuyama, F. (2003). *Our posthuman future: Consequences of the biotechnology revolution*. Farrar, Straus and Giroux.

Fukuyama, F. (2011). *The origins of political order: From prehuman times to the French Revolution*. Farrar, Straus and Giroux.

Fuller, L. L. (1958). Positivism and Fidelity to Law: A Reply to Professor Hart. *Harvard Law Review*, *71*(4), 630-672. https://doi.org/10.2307/1338226

Fumagalli, M., Giannicola, G., Rosa, M., Marceglia, S., Lucchiari, C., Mrakic-Sposta, S., Servello, D., Pacchetti, C., Porta, M., Sassi, M., Zangaglia, R., Franzini, A., Albanese, A., Romito, L., Piacentini, S., Zago, S., Pravettoni, G., Barbieri, S., & Priori, A. (2011). Conflict-dependent dynamic of subthalamic nucleus oscillations during moral decisions. *Soc Neurosci*, *6*(3), 243-256. https://doi.org/10.1080/17470919.2010.515148

Fumagalli, M., Marceglia, S., Cogiamanian, F., Ardolino, G., Picascia, M., Barbieri, S., Pravettoni, G., Pacchetti, C., & Priori, A. (2015). Ethical safety of deep brain stimulation: A study on moral decision-making in Parkinson's disease. *Parkinsonism Relat Disord*, *21*(7), 709-716. https://doi.org/10.1016/j.parkreldis.2015.04.011

Fuss, J., Auer, M. K., Biedermann, S. V., Briken, P., & Hacke, W. (2015). Deep brain stimulation to reduce sexual drive. *Journal of psychiatry & neuroscience: JPN*, *40*(6), 429-431.

Fuster, J. M. (2001). The prefrontal cortex—an update: time is of the essence. *Neuron*, *30*(2), 319-333.

Gagne, P. (1981). Treatment of sex offenders with medroxyprogesterone acetate. *Am J Psychiatry*, *138*(5), 644-646. https://doi.org/10.1176/ajp.138.5.644

Gallagher, S. (2005). *How the body shapes the mind*. Clarendon Press. Publisher description

Gallagher, S. (2014). The cruel and unusual phenomenology of solitary confinement. *FRONT PSYCHOL*, *5*, 585. https://doi.org/10.3389/fpsyg.2014.00585

Gallagher, S. (2018). Deep Brain Stimulation, Self and Relational Autonomy. *Neuroethics*, *14*(1), 31-43. https://doi.org/10.1007/s12152-018-9355-x

Gallagher, S., Morgan, B., & Rokotnitz, N. (2018). Relational Authenticity. In *Neuroexistentialism*. Oxford University Press. https://doi.org/10.1093/oso/9780190460723.003.0008

Gao, Y., Glenn, A. L., Schug, R. A., Yang, Y., & Raine, A. (2009). The neurobiology of psychopathy: a neurodevelopmental perspective. *Can J Psychiatry*, *54*(12), 813-823. https://doi.org/10.1177/070674370905401204

Gardner, A. B., Krieger, A. M., Vachtsevanos, G., Litt, B., & Kaelbing, L. P. (2006). One-class novelty detection for seizure analysis from intracranial EEG. *Journal of Machine Learning Research*, *7*(6).

Gardner, J. (1997). Gist of Excuses, The. *Buff. Crim. L. Rev.*, *1*, 575.

Gardner, J. (2004). The wrongdoing that gets results. *Philosophical Perspectives*, *18*, 53-88.

Garland-Thomson, R. (2012). The case for conserving disability. *J Bioeth Inq*, *9*(3), 339-355. https://doi.org/10.1007/s11673-012-9380-0

Garland, D. (2001). Introduction: The meaning of mass imprisonment. In: Sage Publications.

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, *123*(7), 1293-1326. https://doi.org/10.1093/brain/123.7.1293

Genschow, O., Noll, T., Wänke, M., & Gersbach, R. (2014). Does Baker-Miller pink reduce aggression in prison detention cells? A critical empirical examination. *Psychology, Crime & Law*, *21*(5), 482-489. https://doi.org/10.1080/1068316x.2014.989172

Gert, B. (1998). *Morality: Its nature and justification*. Oxford University Press on Demand.

Gesch, C. B., Hammond, S. M., Hampson, S. E., Eves, A., & Crowder, M. J. (2002). Influence of supplementary vitamins, minerals and essential fatty acids on the

antisocial behaviour of young adult prisoners. Randomised, placebo-controlled trial. *Br J Psychiatry*, *181*(1), 22-28. https://doi.org/10.1192/bjp.181.1.22

Gewirth, A. (1978). *Reason and morality*. Chicago : University of Chicago Press.

Gewirth, A. (1985). Why there are human rights. *Social Theory and Practice*, *11*(2), 235-248.

Gewirth, A. (1998). The community of rights. In *Applied ethics in a troubled world* (pp. 225-235). Springer.

Gilabert, P. (2010). The importance of linkage arguments for the theory and practice of human rights: A response to James Nickel. *Hum. Rts. Q.*, *32*, 425-438.

Gilabert, P. (2019). *Human dignity and human rights*. Oxford University Press, USA.

Gilbert, F., & Dodds, S. (2020). Is There Anything Wrong With Using AI Implantable Brain Devices to Prevent Convicted Offenders from Reoffending? In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity*. Oxford University Press.

Gilbert, F., & Lancelot, M. (2021). Incoming ethical issues for deep brain stimulation: when long-term treatment leads to a 'new form of the disease'. *J Med Ethics*, *47*(1), 20-25. https://doi.org/10.1136/medethics-2019-106052

Gillespie, S. M., Brzozowski, A., & Mitchell, I. J. (2017). Self-regulation and aggressive antisocial behaviour: insights from amygdala-prefrontal and heart-brain interactions. *Psychology, Crime & Law*, *24*(3), 243-257. https://doi.org/10.1080/1068316x.2017.1414816

Gillespie, W. (2002). *Prisonization: Individual and Institutional Factors Affecting Inmate Conduct (Criminal Justice (LFB Scholarly Publishing LLC))*. LFB Scholarly Publisishing LLC.

Gilligan, J. (2000). Punishment and Violence: Is the Criminal Law Based on One Huge Mistake? *SOC RES*, *67*(3), 745-772.

Gillon, R. (2003). Ethics needs principles—four can encompass the rest—and respect for autonomy should be "first among equals". *Journal of medical ethics*, *29*(5), 307-312.

Gilmore, S. (2016). 52 months of torture and zero answers (solitary confinement of Adam Capay). *Maclean's*, *129*(45), 20-20.

Giltay, E. J., & Gooren, L. J. (2009). Potential side effects of androgen deprivation treatment in sex offenders. *J Am Acad Psychiatry Law*, *37*(1), 53-58. https://www.ncbi.nlm.nih.gov/pubmed/19297634

Ginsberg, Y., & Lindefors, N. (2012). Methylphenidate treatment of adult male prison inmates with attention-deficit hyperactivity disorder: randomised double-blind placebo-controlled trial with open-label extension. *Br J Psychiatry*, *200*(1), 68-73. https://doi.org/10.1192/bjp.bp.111.092940

Gintis, H. (2000). *Game theory evolving: A problem-centered introduction to modeling strategic behavior*. Princeton university press.

Glannon, W. (2007). *Bioethics and the brain*. Oxford University Press.

Glannon, W. (2009). Our brains are not us. *BIOETHICS*, *23*(6), 321-329. https://doi.org/10.1111/j.1467-8519.2009.01727.x

Glannon, W. (2014). Intervening in the psychopath's brain. *Theor Med Bioeth*, *35*(1), 43-57. https://doi.org/10.1007/s11017-013-9275-z

Glannon, W. (2015). Neuromodulation and the mind-brain relation. *Front Integr Neurosci*, *9*, 22. https://doi.org/10.3389/fnint.2015.00022

Glannon, W. (2018a). Brain Implants. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Glannon, W. (2018b). *Genes and Future People* (1 ed.). Boulder: Routledge. https://doi.org/10.4324/9780429500237

Glannon, W. (2019a). Agency, Identity and Dementia. In (pp. 50-83).

Glannon, W. (2019b). *The neuroethics of memory : from total recall to oblivion*. Cambridge, United Kingdom : Cambridge University Press.

Glannon, W. (2020). Neuroprosthetics, Behavior Control, and Criminal Responsibility. In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity*. Oxford University Press.

Glendinning, C. (1990). Notes toward a neo-Luddite manifesto. *Utne Reader*, *38*(1), 50-53.

Glendon, M. A. (2002). *A world made new: Eleanor Roosevelt and the Universal Declaration of Human Rights*. Random House Trade Paperbacks.

Glickstein, M. (2014). *Neuroscience: A Historical Introduction*. Cambridge: MIT Press.

Glimcher, P. W., & Fehr, E. (2013). *Neuroeconomics: Decision making and the brain*. Academic Press.

Goddard, E. (2017). Deep Brain Stimulation Through the "Lens of Agency": Clarifying Threats to Personal Identity from Neurological Intervention. *Neuroethics*, *10*(3), 325-335. https://doi.org/10.1007/s12152-016-9297-0

Godfrey-Smith, P. (1998). *Complexity and the Function of Mind in Nature*. Cambridge University Press.

Goering, S. (2018). Thinking Differently. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics* (pp. 37-50). Routledge.

Goering, S., Brown, T., & Klein, E. (2021). Neurotechnology ethics and relational agency. *PHILOS COMPASS*, *16*(4). https://doi.org/10.1111/phc3.12734

Goering, S., Klein, E., Dougherty, D. D., & Widge, A. S. (2017). Staying in the Loop: Relational Agency and Identity in Next-Generation DBS for Psychiatry. *AJOB neuroscience*, *8*(2), 59-70. https://doi.org/10.1080/21507740.2017.1320320

Golan, D. E., & Tashjia, A. H. (2004). *Principles of Pharmacology: The Pathophysiologic Basis of Drug Therapy* (Vol. 28). Ringgold, Inc.

Goldberg, E. (2001). *The executive brain: Frontal lobes and the civilized mind*. Oxford University Press, USA.

Goldman, A. H. (1979). The paradox of punishment. *Philosophy & Public Affairs*, 42-58.

Goldman, A. H. (1982). Toward a new theory of punishment. *Law and Philosophy*, *1*(1), 57-76.

Gordon, M. S., Kinlock, T. W., Couvillion, K. A., Schwartz, R. P., & O'Grady, K. (2012). A Randomized Clinical Trial of Methadone Maintenance for Prisoners: Prediction of Treatment Entry and Completion in Prison. *J Offender Rehabil*, *51*(4), 222-238. https://doi.org/10.1080/10509674.2011.641075

Grasswick, L. J., & Bradford, J. M. (2003). Osteoporosis associated with the treatment of paraphilias: a clinical review of seven case reports. *J Forensic Sci*, *48*(4), 849-855. https://www.ncbi.nlm.nih.gov/pubmed/12877306

Grau, C., Ginhoux, R., Riera, A., Nguyen, T. L., Chauvat, H., Berg, M., Amengual, J. L., Pascual-Leone, A., & Ruffini, G. (2014). Conscious brain-to-brain communication in humans using non-invasive technologies. *PLOS ONE*, *9*(8), e105225. https://doi.org/10.1371/journal.pone.0105225

Greely, H. (2008). Neuroscience and Criminal Justice: Not Responsibility but Treatment. *Kansas Law Review*, *56*(5), 1103. https://doi.org/10.17161/1808.20016

Greely, H., Sahakian, B., Harris, J., Kessler, R. C., Gazzaniga, M., Campbell, P., & Farah, M. J. (2008). Towards responsible use of cognitive-enhancing drugs by the healthy. *Nature*, *456*(7223), 702-705. https://doi.org/10.1038/456702a

Greely, H. T. (2012). Direct brain interventions to "treat" disfavored human behaviors: ethical and social issues. *Clin Pharmacol Ther*, *91*(2), 163-165. https://doi.org/10.1038/clpt.2011.292

Green, L., & Adams, T. (2019). Legal Positivism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Vol. (Winter 2019 Edition)).

Green, R. M. (2010). The President's Council on Bioethics-Requiescat in Pace. *Journal of Religious Ethics*, *38*(2), 197-218. https://doi.org/10.1111/j.1467-9795.2010.00426.x

Green, T. H. (1986). *Lectures on the principles of political obligation and other writings*. Cambridge University Press.

Green, T. H. ([1883] 1990). *Prolegomena to ethics* (A. C. Bradle, Ed.). Clarendon Press.

Green, W. (1986). Depo-Provera, castration, and the probation of rape offenders: Statutory and constitutional issues. *U. Dayton L. Rev.*, *12*, 1.

Greene, J. (2003). From neural 'is' to moral 'ought': what are the moral implications of neuroscientific moral psychology? *NAT REV NEUROSCI*, *4*(10), 846-849. https://doi.org/10.1038/nrn1224

Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philos Trans R Soc Lond B Biol Sci*, *359*(1451), 1775-1785. https://doi.org/10.1098/rstb.2004.1546

Greene, J. D. (2008). The secret joke of Kant's soul. *Moral psychology*, *3*, 35-79.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin Press.

Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*(12), 517-523.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144-1154. https://doi.org/10.1016/j.cognition.2007.11.004

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *NEURON*, *44*(2), 389-400.

Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proc Natl Acad Sci U S A*, *106*(30), 12506-12511. https://doi.org/10.1073/pnas.0900152106

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108. https://doi.org/10.1126/science.1062872

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychol Rev*, *109*(1), 3-25. https://doi.org/10.1037/0033-295x.109.1.3

Griffin, J. (2009). *On Human Rights*. Oxford University Press.

Groll, D., & Lott, M. (2015). Is There a Role for 'Human Nature' in Debates About Human Enhancement? *Philosophy*, *90*(4), 623-651. https://doi.org/10.1017/s0031819115000376

Grossman, N., Bono, D., Dedic, N., Kodandaramaiah, S. B., Rudenko, A., Suk, H.-J., Cassara, A. M., Neufeld, E., Kuster, N., Tsai, L.-H., Pascual-Leone, A., & Boyden, E. S. (2017). Noninvasive Deep Brain Stimulation via Temporally Interfering

Electric Fields. *Cell*, *169*(6), 1029-1041.e1016. https://doi.org/10.1016/j.cell.2017.05.024

Guastello, S. J., Koopmans, M., & Pincus, D. (2008). *Chaos and complexity in psychology: The theory of nonlinear dynamical systems*. Cambridge University Press.

Guskjolen, A., Kenney, J. W., de la Parra, J., Yeung, B. A., Josselyn, S. A., & Frankland, P. W. (2018). Recovery of "Lost" Infant Memories in Mice. *CURR BIOL*, *28*(14), 2283-2290 e2283. https://doi.org/10.1016/j.cub.2018.05.059

Habermas, J. (1984). *The Theory of Communicative Action: Volume 1: Reason and the Rationalization of Society* (T. McCarty, Trans.). Polity.

Habermas, J. (1990). *Moral consciousness and communicative action*. MIT press.

Habermas, J. (1996a). *Between facts and norms: contributions to a discourse theory of law and democracy*. Polity Press.

Habermas, J. (1996b). Between facts and norms: Contributions to a discourse theory of law and democracy, Translated by W Rehg. *Cabridge: Polity*.

Habermas, J. (2003). *The future of human nature* (H. Beister, M. Pensky, & W. Rehg, Trans.). Polity Press.

Habermas, J. (2018). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. John Wiley & Sons.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychol Rev*, *108*(4), 814-834. https://doi.org/10.1037/0033-295x.108.4.814

Haidt, J., & Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research*, *20*(1), 98-116. https://doi.org/10.1007/s11211-007-0034-z

Haji, I. (2009). *Incompatibilism's allure : principal arguments for incompatibilism*. Peterborough, Ont. : Broadview Press.

Haji, I. (2016). *Luck's Mischief*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190260774.001.0001

Haley, H. (2020). The Opioid Crisis as Health Crisis, Not Criminal Crisis: Implications for the Criminal Justice System. *Dalhousie law journal*, *43*(1), 1-34.

Hall, W., & Carter, A. (2013). How may neuroscience affect the way that the criminal courts deal with addicted offenders? . In N. A. Vincent (Ed.), *Neuroscience and legal responsibility*. Oxford University Press.

Hall, W., Ward, J., & Mattick, R. (1993). Methadone maintenance treatment in prisons: The New South Wales Experience. *Drug Alcohol Rev*, *12*(2), 193-203. https://doi.org/10.1080/09595239300185631

Hampson, R. E., Song, D., Opris, I., Santos, L. M., Shin, D. C., Gerhardt, G. A., Marmarelis, V. Z., Berger, T. W., & Deadwyler, S. A. (2013). Facilitation of memory encoding in primate hippocampus by a neuroprosthesis that promotes task-specific neural firing. *Journal of Neural Engineering*, *10*(6), 066013.

Han, S., Mao, L., Gu, X., Zhu, Y., Ge, J., & Ma, Y. (2008). Neural consequences of religious belief on self-referential processing. *Soc Neurosci*, *3*(1), 1-15. https://doi.org/10.1080/17470910701469681

Haney, C., Banks, C., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *The Sociology of Corrections (New York: Wiley, 1977)*, 65-92.

Haney, C., & Zimbardo, P. (1998). The past and future of U.S. prison policy. Twenty-five years after the Stanford prison experiment. *AM PSYCHOL*, *53*(7), 709-727. https://doi.org/10.1037//0003-066x.53.7.709

Hansen, P. G., & Jespersen, A. M. (2013). Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, *4*(1), 3-28.

Hardcastle, V. G. (2018). My Brain Made Me Do It? In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Hardcastle, V. G. (2020). Diversion Courts, Traumatic Brain Injury, and American Vets. In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity*. Oxford University Press.

Hardin, R. (2002). *Trust and trustworthiness*. Russell Sage Foundation.

Hare, R. D. (1991). *The Hare psychopathy checklist-revised: Manual*. Multi-Health Systems, Incorporated.

Hare, R. D. (2003). *The Hare Psychopathy Checklist-Revised* (2nd ed ed.). Multi-Health Systems.

Harenski, C. L., & Hamann, S. (2006). Neural correlates of regulating negative emotions related to moral violations. *neuroimage*, *30*(1), 313-324. https://doi.org/10.1016/j.neuroimage.2005.09.034

Harris, G. T., Rice, M. E., & Cormier, C. A. (1991). Psychopathy and violent recidivism. *Law and Human Behavior*, *15*(6), 625-637.

Harris, J. (2011). Moral enhancement and freedom. *BIOETHICS*, *25*(2), 102-111. https://doi.org/10.1111/j.1467-8519.2010.01854.x

Harris, J. (2012). What it's like to be good. *Camb Q Healthc Ethics*, *21*(3), 293-305. https://doi.org/10.1017/S0963180111000867

Harris, J. (2013a). 'Ethics is for bad guys!' Putting the 'moral' into moral enhancement. *BIOETHICS*, *27*(3), 169-173. https://doi.org/10.1111/j.1467-8519.2011.01946.x

Harris, J. (2013b). Moral progress and moral enhancement. *BIOETHICS*, *27*(5), 285-290.

Harris, J. (2014a). ". . . How narrow the strait!". The God machine and the spirit of liberty. *Camb Q Healthc Ethics*, *23*(3), 247-260. https://doi.org/10.1017/S0963180113000856

Harris, J. (2014b). Taking liberties with free fall. *J Med Ethics*, *40*(6), 371-374. https://doi.org/10.1136/medethics-2012-101092

Harris, J. (2016). *How to be good : the possibility of moral enhancement*. Oxford : Oxford University Press; First edition.

Harris, J. (2016). Moral Blindness - The Gift of the God Machine. *Neuroethics*, *9*(3), 269-273. https://doi.org/10.1007/s12152-016-9272-9

Harris, J. C. (2003). Social neuroscience, empathy, brain integration, and neurodevelopmental disorders. *Physiol Behav*, *79*(3), 525-531. https://doi.org/10.1016/s0031-9384(03)00158-6

Hart, H., & Rubia, K. (2012). Neuroimaging of child abuse: a critical review. *FRONT HUM NEUROSCI*, *6*(2012), 52. https://doi.org/10.3389/fnhum.2012.00052

Hart, H. L. A. (1955). Are there any natural rights? *The Philosophical Review*, *64*(2), 175-191.

Hart, H. L. A. (1961 [2012]). *The Concept of Law* (Third Edition ed.). Clarendon Press.

Hart, H. L. A. (1963). *Law, liberty, and morality*. Stanford University Press.

Hart, H. L. A. (1979). Between utility and rights. *Columbia Law Review*, *79*(5), 828-846.

Hart, H. M. (1958). The aims of the criminal law. *Law and Contemporary problems*, *23*(3), 401-441.

Haslam, N. (2006). Dehumanization: an integrative review. *Pers Soc Psychol Rev*, *10*(3), 252-264. https://doi.org/10.1207/s15327957pspr1003_4

Hauskeller, M. (2011). Human Enhancement and the Giftedness of Life. *Philosophical Papers*, *40*(1), 55-79. https://doi.org/10.1080/05568641.2011.560027

Hauskeller, M. (2017). Is It Desirable to Be Able to Do the Undesirable? Moral Bioenhancement and the Little Alex Problem. *Camb Q Healthc Ethics*, *26*(3), 365-376. https://doi.org/10.1017/S096318011600102X

Hedrich, D., Alves, P., Farrell, M., Stover, H., Moller, L., & Mayet, S. (2012). The effectiveness of opioid maintenance treatment in prison settings: a systematic review. *Addiction*, *107*(3), 501-517. https://doi.org/10.1111/j.1360-0443.2011.03676.x

Heersmink, R. (2014). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences*, *14*(3), 577-598. https://doi.org/10.1007/s11097-014-9355-1

Heersmink, R. (2016). Distributed selves: personal identity and extended memory systems. *Synthese*, *194*(8), 3135-3151. https://doi.org/10.1007/s11229-016-1102-4

Heffernan, W. C., & Kleinig, J. (2000). *From social justice to criminal justice: Poverty and the administration of criminal law*. Oxford University Press on Demand.

Hegel, G. W. F. (1956). *The Philosophy of History*. Dover.

Hegel, G. W. F. (1987). *Introduction to the Lectures on the History of Philosophy* (T. M. Knox & A. V. Miller, Trans.). Oxford University Press.

Hegel, G. W. F. (2017 [1807]). *The phenomenology of spirit* (T. P. Pinkard, Trans.). Cambridge University Press.

Hegel, G. W. F. ([1821] 1991). *Hegel: Elements of the philosophy of right* (A. W. Wood & R. E. Nisbett, Trans.). Cambridge University Press.

Heidegger, M. (1962 [1927]). *Being and Time* (M. J. & R. E., Trans.). Harper & Row.

Heinrichs, J.-H. (2018). Neuroethics, Cognitive Technologies and the Extended Mind Perspective. *Neuroethics*, *14*(1), 59-72. https://doi.org/10.1007/s12152-018-9365-8

Heinrichs, J.-H., & Stake, M. (2018). Enhancement: Consequentialist arguments. *Zeitschrift für Ethik und Moralphilosophie*, *1*(2), 321-342.

Heinrichs, J.-H., & Stake, M. (2019). Human Enhancement: Arguments from Virtue Ethics. *Zeitschrift für Ethik und Moralphilosophie*, *2*(2), 355-373.

Heldke, L., & Thomsen, J. (2014). Two concepts of authenticity. *Social Philosophy Today*.

Hendershot, A. (2020). Solving the Fentanyl Problem beyond the Border: A Call for an International Solution. *Penn St. JL & Int'l Aff.*, *9*, 216.

Henrichson, C., & Delaney, R. (2012). The Price of Prisons: What Incarceration Costs Taxpayers. *Federal sentencing reporter*, *25*(1), 68-80. https://doi.org/10.1525/fsr.2012.25.1.68

Herman, J. P., McKlveen, J. M., Ghosal, S., Kopp, B., Wulsin, A., Makinson, R., Scheimann, J., & Myers, B. (2016). Regulation of the Hypothalamic-Pituitary-Adrenocortical Stress Response. *Compr Physiol*, *6*(2), 603-621. https://doi.org/10.1002/cphy.c150015

Hertwig, R., & Grune-Yanoff, T. (2017). Nudging and Boosting: Steering or Empowering Good Decisions. *Perspect Psychol Sci*, *12*(6), 973-986. https://doi.org/10.1177/1745691617702496

Hess, U., & Blairy, S. (2001). Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International journal of psychophysiology*, *40*(2), 129-141.

Heuer, U. (2022). Introduction. In U. Heuer (Ed.), *The Roots of Normativity* (pp. 1-18). Oxford University Press. https://doi.org/10.1093/oso/9780192847003.003.0001

Heylighen, F., Cilliers, P., & Gershenson, C. (2006). Complexity and philosophy. *arXiv preprint cs/0604072*.

Hill, A. T., Fitzgerald, P. B., & Hoy, K. E. (2016). Effects of Anodal Transcranial Direct Current Stimulation on Working Memory: A Systematic Review and Meta-Analysis

of Findings From Healthy and Neuropsychiatric Populations. *Brain Stimul*, *9*(2), 197-208. https://doi.org/10.1016/j.brs.2015.10.006

Hillis, A. E. (2014). Inability to empathize: brain lesions that disrupt sharing and understanding another's emotions. *Brain*, *137*(4), 981-997.

Hirsch, A. v. (1996). *Censure and Sanctions*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198262411.001.0001

Hobbes, T. ([1651] 1980). *Leviathan*

Hodges, A. (1983). Alan Turing: The Enigma. In. Princeton: Princeton University Press.

Hoekema, D. A. (1986). *Rights and wrongs: Coercion, punishment, and the state*. Susquehanna University Press.

Hof, W. (2020). *The Wim Hof Method: Activate Your Full Human Potential*. Sounds True.

Hohfeld, W. N. (1913). Some fundamental legal conceptions as applied in judicial reasoning. *Yale Lj*, *23*, 16.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *7*(2), 174-196.

Holm, S., & McNamee, M. (2011). Physical Enhancement. In *Enhancing Human Capacities* (pp. 291-303). https://doi.org/10.1002/9781444393552.ch21

Holmen, S. J. (2020). Respect, Punishment and Mandatory Neurointerventions. *Neuroethics*, *14*(2), 167-176. https://doi.org/10.1007/s12152-020-09434-8

Holroyd, J. (2010). Punishment and justice. *Social Theory and Practice*, *36*(1), 78-111.

Holroyd, J., Scaife, R., & Stafford, T. (2017). Responsibility for implicit bias. *Philosophy Compass*, *12*(3), e12410-n/a. https://doi.org/10.1111/phc3.12410

Homan, J. (2011). *Charlatans, seekers, and shamans: The ayahuasca boom in western Peruvian Amazonia* University of Kansas].

Hoogman, M., Bralten, J., Hibar, D. P., Mennes, M., Zwiers, M. P., Schweren, L. S. J., van Hulzen, K. J. E., Medland, S. E., Shumskaya, E., Jahanshad, N., Zeeuw, P., Szekely, E., Sudre, G., Wolfers, T., Onnink, A. M. H., Dammers, J. T., Mostert, J. C., Vives-Gilabert, Y., Kohls, G., . . . Franke, B. (2017). Subcortical brain volume differences in participants with attention deficit hyperactivity disorder in children

and adults: a cross-sectional mega-analysis. *Lancet Psychiatry*, *4*(4), 310-319. https://doi.org/10.1016/S2215-0366(17)30049-4

Hornblum, A. M. (1997). They were cheap and available: prisoners as research subjects in twentieth century America. *BMJ*, *315*(7120), 1437-1441.

Hosseini, S. H., Mano, Y., Rostami, M., Takahashi, M., Sugiura, M., & Kawashima, R. (2011). Decoding what one likes or dislikes from single-trial fNIRS measurements. *Neuroreport*, *22*(6), 269-273.

Hough, M., Clancy, A., McSweeney, T., & Turnbull, P. J. (2003). *The impact of drug treatment and testing orders on offending: Two-year reconviction results*. Home Office. Research, Development and Statistics Directorate.

Hrymak, H. (2018). A Bad Deal: British Columbia's Emphasis on Deterrence and Increasing Prison Sentences for Street-Level Fentanyl Traffickers. *Manitoba law journal (1966)*, *41*(4), 149.

Hsu, C. W., Begliomini, C., Dall'Acqua, T., & Ganis, G. (2019). The effect of mental countermeasures on neuroimaging-based concealed information tests. *HUM BRAIN MAPP*, *40*(10), 2899-2916. https://doi.org/10.1002/hbm.24567

Huang, P. H. (2020). Who's afraid of perfectionist moral enhancement? A reply to Sparrow. *BIOETHICS*, *34*(8), 865-871. https://doi.org/10.1111/bioe.12751

Huang, Z., Liu, X., Mashour, G. A., & Hudetz, A. G. (2018). Timescales of Intrinsic BOLD Signal Dynamics and Functional Connectivity in Pharmacologic and Neuropathologic States of Unconsciousness. *J NEUROSCI*, *38*(9), 2304-2317. https://doi.org/10.1523/JNEUROSCI.2545-17.2018

Hume, D., & Millican, P. J. R. (1748/2007). *An enquiry concerning human understanding*. Oxford University Press.

Hurley, S. L. (1998). *Consciousness in action*. Cambridge, Mass. : Harvard University Press.

Husak, D. (2007). *Overcriminalization*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195328714.001.0001

Husak, D. (2008). *Overcriminalization: The limits of the criminal law*. Oxford University Press.

Husak, D. N. (1992). Why punish the deserving? *Nous*, *26*(4), 447-464.

Hutchins, E. (1995). *Cognition in the Wild*. MIT press.

Huxley, A. (2021 [1932]). *A Brave New World*. Good Press.

Hysek, C. M., Schmid, Y., Simmler, L. D., Domes, G., Heinrichs, M., Eisenegger, C., Preller, K. H., Quednow, B. B., & Liechti, M. E. (2014). MDMA enhances emotional empathy and prosocial behavior. *Soc Cogn Affect Neurosci*, *9*(11), 1645-1652. https://doi.org/10.1093/scan/nst161

Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sci Soc Policy*, *13*(1), 5. https://doi.org/10.1186/s40504-017-0050-1

Illes, J., & Racine, E. (2005). Imaging or imagining? A neuroethics challenge informed by genetics. *Am J Bioeth*, *5*(2), 5-18. https://doi.org/10.1080/15265160590923358

Illia, L., Colleoni, E., & Zyglidopoulos, S. (2023). Ethical implications of text generation in the age of artificial intelligence. *Business Ethics, the Environment & Responsibility*, *32*(1), 201-210.

Inglese, S., & Lavazza, A. (2021). What Should We Do With People Who Cannot or Do Not Want to Be Protected From Neurotechnological Threats? [Brief article]. *Frontiers in Human Neuroscience*, NA.

Ingvar, D. H. (1979). "Hyperfrontal" distribution of the cerebral grey matter flow in resting wakefulness; on the functional anatomy of the conscious state. *Acta Neurol Scand*, *60*(1), 12-25. https://doi.org/10.1111/j.1600-0404.1979.tb02947.x

Ishay, M. (2008). *The history of human rights: From ancient times to the globalization era*. Univ of California Press.

Ivison, D. (2007). *Rights*. McGill-Queen's University Press.

Jacobs, J. (2016). From Bad to Worse: Crime, Incarceration, and the Self-Wounding Society. In S. Farrall, B. Goldson, I. Loader, & A. Dockley (Eds.), *Justice and Penal Reform*. Routledge.

James, D. J., & Glaze, L. E. (2006). Mental health problems of prison and jail inmates.

Jansen, J. M., Daams, J. G., Koeter, M. W., Veltman, D. J., van den Brink, W., & Goudriaan, A. E. (2013). Effects of non-invasive neurostimulation on craving: a

meta-analysis. *Neurosci Biobehav Rev*, *37*(10 Pt 2), 2472-2480. https://doi.org/10.1016/j.neubiorev.2013.07.009

Jean-Richard-Dit-Bressel, P., Killcross, S., & McNally, G. P. (2018). Behavioral and neurobiological mechanisms of punishment: implications for psychiatric disorders. *Neuropsychopharmacology*, *43*(8), 1639-1650. https://doi.org/10.1038/s41386-018-0047-3

Jenkins, A. (2020). *The portfolio diet of foods to lower cholesterol and reduce cardiovascular disease : an evidence based approach for plant food consumption*. London, England : Academic Press.

Jeurissen, D., Sack, A. T., Roebroeck, A., Russ, B. E., & Pascual-Leone, A. (2014). TMS affects moral judgment, showing the role of DLPFC and TPJ in cognitive and emotional processing. *Frontiers in neuroscience*, *8*(8), 18. https://doi.org/10.3389/fnins.2014.00018

Jiang, L., Stocco, A., Losey, D. M., Abernethy, J. A., Prat, C. S., & Rao, R. P. N. (2019). BrainNet: A Multi-Person Brain-to-Brain Interface for Direct Collaboration Between Brains. *Sci Rep*, *9*(1), 6115. https://doi.org/10.1038/s41598-019-41895-7

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816-821. https://doi.org/10.1126/science.1225829

Johnson, J. (2014). *American lobotomy: A rhetorical history*. University of Michigan Press.

Joseph, R. S. (2008). Functional MRI Lie Detection: Too Good to be True? *J AM ACAD PSYCHIATRY*, *36*(4), 491-498.

Juneau, M. (2018). *Cardiovascular health : living your best with a healthy heart*. Toronto : Dundurn.

Jung, C. G. (1959). *The archetypes and the collective unconscious*. Routledge.

Kahane, G., Pugh, J., & Savulescu, J. (2017). Bioconservatism, Partiality, and the Human-Nature Objection to Enhancement. *The Monist*, *99*(4), 406-422. https://doi.org/10.1093/monist/onw013

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kalis, A., Mojzisch, A., Schweizer, T. S., & Kaiser, S. (2008). Weakness of will, akrasia, and the neuropsychiatry of decision making: an interdisciplinary perspective. *Cogn Affect Behav Neurosci*, *8*(4), 402-417. https://doi.org/10.3758/CABN.8.4.402

Kamm, F. M. (2007). *Intricate ethics : rights, responsibilities, and permissible harm*. New York : Oxford University Press.

Kandel, E. R. (2013). *Principles of neural science*. New York : McGraw-Hill Medical; 5th ed. / edited by Eric R. Kandel ... [et al.] ; art editor, Sarah Mack.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., & Mack, S. (2000). *Principles of neural science* (Vol. 4). McGraw-hill New York.

Kane, R. (1998). *The significance of free will*. Oxford University Press on Demand.

Kant, I. (1797). *The Metaphysics of Morals, in Immanuel Kant: Practical Philosophy* (M. J. Gregor, Trans.). Cambridge University Press.

Kant, I. (1998). *Kant: Religion within the boundaries of mere reason: And other writings*. Cambridge University Press.

Kant, I. (1999 (1781)). *Critique of pure reason* (P. Guyer & A. W. Wood, Eds.). Cambridge University Press.

Kant, I. ([1785] 2002). *Groundwork for the Metaphysics of Morals* (J. B. Schneewind, Trans.). Yale University Press.

Kass, L. (2003). *Beyond therapy: biotechnology and the pursuit of happiness*. Executive Office of the President.

Kaufman, W. (2008). The rise and fall of the mixed theory of punishment. *International Journal of Applied Philosophy*, *22*(1), 37-57.

Keating, N. L., O'Malley, A. J., & Smith, M. R. (2006). Diabetes and cardiovascular disease during androgen deprivation therapy for prostate cancer. *J Clin Oncol*, *24*(27), 4448-4456. https://doi.org/10.1200/JCO.2006.06.2497

Kellmeyer, P., Cochrane, T., Müller, O., Mitchell, C., Ball, T., Fins, J. J., & Biller-Andorno, N. (2017). The Effects of Closed-Loop Medical Devices on the

Autonomy and Accountability of Persons and Systems-CORRIGENDUM. *Camb Q Healthc Ethics*, *26*(1), 180-180. https://doi.org/10.1017/S0963180116000967

Kelso, J. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. MIT press.

Kent, L., & Wittmann, M. (2021). Special Issue: Consciousness science and its theories Time consciousness: the missing link in theories of consciousness. *Neurosci Conscious*, *2021*(1), niab011. https://doi.org/10.1093/nc/niab011

Kesey, K. (2007). *One Flew Over the Cuckoo's Nest:(Penguin Classics Deluxe Edition)*. Penguin.

Kessler, R. C., Davis, C. G., & Kendler, K. S. (1997). Childhood adversity and adult psychiatric disorder in the US National Comorbidity Survey. *Psychol Med*, *27*(5), 1101-1119. https://doi.org/10.1017/s0033291797005588

Khantzian, E. J. (1997). The self-medication hypothesis of substance use disorders: a reconsideration and recent applications. *Harv Rev Psychiatry*, *4*(5), 231-244. https://doi.org/10.3109/10673229709030550

Kiefer, M., Adams, S. C., & Zovko, M. (2012). Attentional sensitization of unconscious visual processing: Top-down influences on masked priming. *Adv Cogn Psychol*, *8*(1), 50-61. https://doi.org/10.2478/v10053-008-0102-4

Kiehl, K. A. (2006). A cognitive neuroscience perspective on psychopathy: evidence for paralimbic system dysfunction. *Psychiatry Res*, *142*(2-3), 107-128. https://doi.org/10.1016/j.psychres.2005.09.013

Kiehl, K. A., & Hoffman, M. B. (2011). The criminal psychopath: History, neuroscience, treatment, and economics. *Jurimetrics*, *51*, 355.

Kim, J. (2007). *Physicalism, or something near enough*. Princeton University Press.

Kim, J. (2018). *Philosophy of mind*. Routledge.

Kim, J., Sosa, E., & Korman, D. Z. (2012). *Metaphysics : an anthology* (2nd ed.). Wiley-Blackwell. Cover image http://catalogimages.wiley.com/images/db/jimages/9781444331011.jpg

King, M. (2014). Two faces of desert. *Philosophical Studies*, *169*(3), 401-424.

King, S. (2010). *Different Seasons*. Hodder & Stoughton. http://books.google.ca/books?id=XSIUoqwXhH0C

Kishiyama, M. M., Boyce, W. T., Jimenez, A. M., Perry, L. M., & Knight, R. T. (2009). Socioeconomic disparities affect prefrontal function in children. *J Cogn Neurosci*, *21*(6), 1106-1115. https://doi.org/10.1162/jocn.2009.21101

Kitajka, K., Sinclair, A. J., Weisinger, R. S., Weisinger, H. S., Mathai, M., Jayasooriya, A. P., Halver, J. E., & Puskás, L. G. (2004). Effects of dietary omega-3 polyunsaturated fatty acids on brain gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(30), 10931-10936. https://doi.org/10.1073/pnas.0402342101

Klein, E., Goering, S., Gagne, J., Shea, C. V., Franklin, R., Zorowitz, S., Dougherty, D. D., & Widge, A. S. (2016). Brain-computer interface-based control of closed-loop brain stimulation: attitudes and ethical considerations. *Brain-Computer Interfaces*, *3*(3), 140-148. https://doi.org/10.1080/2326263x.2016.1207497

Kleinig, J. (2012). *Ethics and Criminal Justice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511806155

Kline, N. S. (1959). Psychopharmaceuticals: effects and side effects. *Bull World Health Organ*, *21*(4-5), 397-410. https://www.ncbi.nlm.nih.gov/pubmed/14409889

Klinger, E. (1971). Structure and functions of fantasy.

Knack, N., Chandler, J. A., & Fedoroff, J. P. (2020). A qualitative study of forensic patients' perceptions of quasi-coercive offers of biological treatment. *Behavioral Sciences & the Law*, *38*(2), 135-151. https://doi.org/10.1002/bsl.2449

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908-911. https://doi.org/10.1038/nature05631

Kolk, B. A. v. d., & Fisler, R. E. (1994). Childhood abuse and neglect and loss of self-regulation. *Bulletin of the Menninger Clinic*, *58*(2), 145.

Korn, H. A., Johnson, M. A., & Chun, M. M. (2012). Neurolaw: Differential brain activity for black and white faces predicts damage awards in hypothetical employment

discrimination cases. *Soc Neurosci*, *7*(4), 398-409. https://doi.org/10.1080/17470919.2011.631739

Kornfield, J. (2009). *The wise heart: A guide to the universal teachings of Buddhist psychology*. Bantam.

Koroshetz, W. J., Ward, J., & Grady, C. (2020). NeuroEthics and the BRAIN Initiative: Where Are We? Where Are We Going? *AJOB Neurosci*, *11*(3), 140-147. https://doi.org/10.1080/21507740.2020.1778119

Kretschmann, H.-J. (2004). *Cranial neuroimaging and clinical neuroanatomy : atlas of MR imaging and computed tomography*. Stuttgart : Thieme; 3rd ed., rev. and expanded.

Kripke, S. A. (1980). *Naming and necessity*. Cambridge, Mass. : Harvard University Press.

Kuflik, A. (1984). Henry Shue, "Basic Rights: Subsistence, Affluence, and U.S. Foreign Policy" (Book Review). *Ethics*, *94*(2), 319-319.

Kulynych, J. J. (2007). The regulation of MR neuroimaging research: disentangling the Gordian knot. *Am J Law Med*, *33*(2-3), 295-317. https://doi.org/10.1177/009885880703300207

Kutcher, M. R. (2010a). The chemical castration of recidivist sex offenders in Canada: a matter of faith. *Dalhousie law journal*, *33*(2), 193.

Kutcher, M. R. (2010b). The chemical castration of recidivist sex offenders in Canada: a matter of faith. *Dalhousie law journal*, *33*(2), 193-216.

Laertius, D. (1853). *Lives and Opinions of Eminent Philosophers*. Henry G. Bohn.

Lamb, J. W. (1977). On a Proof of Incompatibilism. *The Philosophical Review*, *86*(1), 20-35. https://doi.org/10.2307/2184160

Langan, P. A., & Levin, D. J. (2002). *Recidivism of prisoners released in 1994*. US Department of Justice, Office of Justice Programs, Bureau of Justice ….

Larkin, P. J., Jr. (2013). Public choice theory and overcriminalization. *Harvard journal of law and public policy*, *36*(2), 715.

Laschet, U., & Laschet, L. (1975). Antiandrogens in the treatment of sexual deviations of men. *J Steroid Biochem*, *6*(6), 821-826. https://doi.org/10.1016/0022-4731(75)90310-6

Laub, J. H. (2003). *Shared beginnings, divergent lives : delinquent boys to age 70*. Harvard University Press.

Laureys, S. (2015). *Un si brillant cerveau: les états limites de conscience*. Odile Jacob.

Lavazza, A. (2017). Neurolaw and punishment: a naturalistic and humanitarian view, and its overlooked perils. *Teoria. Rivista di filosofia*, *37*(2), 81-97.

Lavazza, A. (2018). Freedom of Thought and Mental Integrity: The Moral Requirements for Any Neural Prosthesis [Brief article

Report]. *Frontiers in neuroscience*, *12*, 82-82.

Le Texier, T. (2019). Debunking the Stanford Prison Experiment. *AM PSYCHOL*, *74*(7), 823-839. https://doi.org/10.1037/amp0000401

Lebedev, M. (2014). Brain-machine interfaces: an overview. *Translational Neuroscience*, *5*(1), 99-110.

Lebedev, M. A., & Nicolelis, M. A. (2006). Brain-machine interfaces: past, present and future. *Trends Neurosci*, *29*(9), 536-546. https://doi.org/10.1016/j.tins.2006.07.004

Lee, N., Broderick, A. J., & Chamberlain, L. (2007). What is 'neuromarketing'? A discussion and agenda for future research. *International journal of psychophysiology*, *63*(2), 199-204.

Lee, S.-K. (2012). Self-Determination and the Categories of Freedom in Kant's Moral Philosophy. *Kant-Studien*, *103*(3), 337-350. https://doi.org/10.1515/kant-2012-0022

Leech, R., Kamourieh, S., Beckmann, C. F., & Sharp, D. J. (2011). Fractionating the default mode network: distinct contributions of the ventral and dorsal posterior cingulate cortex to cognitive control. *J NEUROSCI*, *31*(9), 3217-3224. https://doi.org/10.1523/JNEUROSCI.5626-10.2011

Leefmann, B., & Hildt, E. (2018). Neuroethics and the Neuroscientific Turn. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Leutgeb, V., Wabnegger, A., Leitner, M., Zussner, T., Scharmuller, W., Klug, D., & Schienle, A. (2016). Altered cerebellar-amygdala connectivity in violent offenders:

A resting-state fMRI study. *Neurosci Lett*, *610*, 160-164. https://doi.org/10.1016/j.neulet.2015.10.063

Levenson, R. W. (1999). The Intrapersonal Functions of Emotion. *Cognition & Emotion*, *13*(5), 481-504. https://doi.org/10.1080/026999399379159

Levy, N. (2007). *Neuroethics: Challenges for the 21st century*. Cambridge University Press.

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. OUP Oxford.

Levy, N. (2017). Nudges in a post-truth world. *J Med Ethics*, *43*(8), 495-500. https://doi.org/10.1136/medethics-2017-104153

Levy, N. (2019). Due deference to denialism: Explaining ordinary people's rejection of established scientific findings. *Synthese*, *196*(1), 313-327.

Levy, N. (2020). Cognitive Enchancement: Defending the Parity Principle. In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity* (pp. 33-72). Oxford University Press.

Levy, N., & McKenna, M. (2009). Recent Work on Free Will and Moral Responsibility. *Philosophy Compass*, *4*(1), 96-133. https://doi.org/10.1111/j.1747-9991.2008.00197.x

Lewens, T. (2012). Human Nature: The Very Idea. *Philosophy & Technology*, *25*(4), 459-474. https://doi.org/10.1007/s13347-012-0063-x

Lewis, C. S. (1953). The humanitarian theory of punishment. *Res Judicatae*, *6*, 224.

Lewis, D. (1970). How to define theoretical terms. *The Journal of philosophy*, *67*(13), 427-446.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, *50*(3), 249-258.

Lewis, D. K. (1966). An argument for the identity theory. *The Journal of philosophy*, *63*(1), 17-25.

Lewis, D. R. (1976). Survival and Identity. In A. l. Rorty (Ed.), *The Identities of persons*. University of California Press.

Lewis, J. (2021). Autonomy and the limits of cognitive enhancement. *BIOETHICS*, *35*(1), 15-22. https://doi.org/10.1111/bioe.12791

Li, G., & Zhang, D. (2016). Brain-Computer Interface Controlled Cyborg: Establishing a Functional Information Transfer Pathway from Human Brain to Cockroach Brain. *PLOS ONE*, *11*(3), e0150667. https://doi.org/10.1371/journal.pone.0150667

Li, G., & Zhang, D. (2017). Brain-Computer Interface Controlling Cyborg: A Functional Brain-to-Brain Interface Between Human and Cockroach. In *Brain-Computer Interface Research* (pp. 71-79). Springer International Publishing. https://doi.org/10.1007/978-3-319-57132-4_6

Liberto, H. (2018). Chemical Castration and the Violation of Sexual Rights. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press.

Lieberman, M. D. (2013). *Social: Why our brains are wired to connect*. Oxford University Press.

Ligthart, S., Douglas, T., Bublitz, J.-C., & Meynen, G. (2019). The Future of Neuroethics and the Relevance of the Law. *AJOB Neurosci*, *10*(3), 120-121. https://doi.org/10.1080/21507740.2019.1632961

Lilienfeld, S. O., Aslinger, E., Marshall, J., & Satel, S. L. (2018). Neurohype. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Lin, D., Boyle, M. P., Dollar, P., Lee, H., Lein, E. S., Perona, P., & Anderson, D. J. (2011). Functional identification of an aggression locus in the mouse hypothalamus. *Nature*, *470*(7333), 221-226. https://doi.org/10.1038/nature09736

Lin, E. C., Escott, E., Alexander, D. A., & Bleicher, A. G. (2008). *Practical Differential Diagnosis for CT and MRI*. New York: Thieme Medical Publishers, Incorporated.

Lin, P., Fang, Z., Liu, J., & Lee, J. H. (2016). Optogenetic Functional MRI. *J Vis Exp*(110). https://doi.org/10.3791/53346

Lipina, S. J. (2014). Biological and sociocultural determinants of neurocognitive development: central aspects of the current scientific agenda. *Bread and brain, education and poverty*, 1-30.

Lippert-Rasmussen, K. (2018). The Self-Ownership Trilemma, Extended Minds, and Neurointerventions. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press.

Lippke, R. (2001). Criminal offenders and right forfeiture. *Journal of social philosophy, 32*(2001), 78-89.

Lippke, R. L. (1998). Arguing against inhumane and degrading punishment. *Criminal Justice Ethics*, *17*(1), 29-41. https://doi.org/10.1080/0731129x.1998.9992045

Liu, Y., Gu, N., Cao, X., Zhu, Y., Wang, J., Smith, R. C., & Li, C. (2021). Effects of transcranial electrical stimulation on working memory in patients with schizophrenia: A systematic review and meta-analysis. *Psychiatry research*, *296*, 113656-113656. https://doi.org/https://doi.org/10.1016/j.psychres.2020.113656

Locke, J. (1824a). *An essay concerning human understanding* (24th ed.). Printed for C. and J. Rivington ; Longman and Co.

Locke, J. (1824b). *Two treatises of government* (New ed.). Printed for C. and J. Rivington.

Locke, J. (1824c). *The works of John Locke, in nine volumes* (12th ed.). Printed for C. and J. Rivington etc.

Locke, J. ([1651] 2016). *Two treatises of government* (L. Ward, Ed.). Focus.

Locke, J. ([1698] 1998). *A letter concerning toleration* (I. Shapiro, Trans.). Yale University Press.

Loewenstein, G., Rick, S., & Cohen, J. D. (2008). Neuroeconomics. *Annu Rev Psychol*, *59*, 647-672. https://doi.org/10.1146/annurev.psych.59.103006.093710

Lu, F. M., & Yuan, Z. (2015). PET/SPECT molecular imaging in clinical neuroscience: recent advances in the investigation of CNS diseases. *Quantitative imaging in medicine and surgery*, *5*(3), 433-447. https://doi.org/10.3978/j.issn.2223-4292.2015.03.16

Luber, B., & Lisanby, S. H. (2014). Enhancement of human cognitive performance using transcranial magnetic stimulation (TMS). *neuroimage*, *85 Pt 3*(0 3), 961-970. https://doi.org/10.1016/j.neuroimage.2013.06.007

Luby, B. J. (1998). The nature of consciousness: Philosophical debates. *Journal of the history of the behavioral sciences*, *34*(4), 433-434. https://doi.org/10.1002/(sici)1520-6696(199823)34:4<433::Aid-jhbs38>3.0.Co;2-v

Lucke, J., & Partridge, B. (2012). Towards a Smart Population: A Public Health Framework for Cognitive Enhancement. *Neuroethics*, *6*(2), 419-427. https://doi.org/10.1007/s12152-012-9167-3

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, *74*(5), 570-581.

Ma, Y., Wang, C., & Han, S. (2011). Neural responses to perceived pain in others predict real-life monetary donations in different socioeconomic contexts. *neuroimage*, *57*(3), 1273-1280. https://doi.org/10.1016/j.neuroimage.2011.05.003

MacIntyre, A. (2007). *After virtue: a study in moral theory*. University of Notre Dame Press.

MacIntyre, A. (2013). *After virtue*. A&C Black.

MacIntyre, A. C. (1981). *After virtue : a study in moral theory*. University of Notre Dame Press.

MacIntyre, A. C. (1984). *After virtue : a study in moral theory* (2nd ed.). University of Notre Dame Press.

Mackenzie, C., & Stoljar, N. (2000). *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. Oxford University Press.

Mackenzie, C., & Walker, M. (2015). Neurotechnologies, personal identity, and the ethics of authenticity. In *Handbook of neuroethics* (pp. 373-392). Springer, Springer Nature.

MacKenzie, D. L. (2006). *What works in corrections: reducing the criminal activities of offenders and deliquents*. Cambridge University Press.

Mackie, J. L. (1977). *Inventing right and wrong*. Penguin.

Mahmoudi, P., Veladi, H., & Pakdel, F. G. (2017). Optogenetics, Tools and Applications in Neurobiology. *Journal of medical signals and sensors*, *7*(2), 71-79. https://www.ncbi.nlm.nih.gov/pubmed/28553579

Maier, K. (2020). Canada's 'Open Prisons': Hybridisation and the Role of Halfway Houses in Penal Scholarship and Practice. *The Howard Journal of Crime and Justice*, *59*(4), 381-399. https://doi.org/10.1111/hojo.12365

Maier, L. J., Ferris, J. A., & Winstock, A. R. (2018). Pharmacological cognitive enhancement among non-ADHD individuals-A cross-sectional study in 15 countries. *Int J Drug Policy*, *58*, 104-112. https://doi.org/10.1016/j.drugpo.2018.05.009

Malakieh, J. (2020). *Adult and youth correctional statistics in Canada, 2018/2019*. Toronto: Statistics Canada

Mangat, R. (2018). Centring the Margins: Reflections on British Columbia Civil Liberties Association and John Howard Society of Canada v. Canada. *BC studies*(197), 123.

Manson, N. C., & O'Neill, O. (2012). *Rethinking Informed Consent in Bioethics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511814600

Marciniak, R., Sheardova, K., Cermakova, P., Hudecek, D., Sumec, R., & Hort, J. (2014). Effect of meditation on cognitive functions in context of aging and neurodegenerative diseases [Review]. *Front Behav Neurosci*, *8*(17), 17. https://doi.org/10.3389/fnbeh.2014.00017

Markel, D. (2012). Retributive justice and the demands of democratic citizenship. *Va. J. Crim. L.*, *1*, 1.

Marko, D., Bahensky, P., Bunc, V., Grosicki, G. J., & Vondrasek, J. D. (2022). Does Wim Hof Method Improve Breathing Economy during Exercise? *J Clin Med*, *11*(8), 2218. https://doi.org/10.3390/jcm11082218

Marlowe, D. B. (2021). Drug Courts. In N. el-Guebaly, G. Carrà, M. Galanter, & A. M. Baldacchino (Eds.), *Textbook of Addiction Treatment* (pp. 1437-1449). Springer International Publishing. https://doi.org/10.1007/978-3-030-36391-8_101

Martin, R. (1993). *A system of rights*. Clarendon Press.

Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776-798. https://doi.org/10.1016/j.neuron.2020.01.026

Maslen, H., Cheeran, B., Pugh, J., Pycroft, L., Boccard, S., Prangnell, S., Green, A. L., FitzGerald, J., Savulescu, J., & Aziz, T. (2018). Unexpected Complications of Novel Deep Brain Stimulation Treatments: Ethical Issues and Clinical Recommendations. *Neuromodulation*, *21*(2), 135-143. https://doi.org/10.1111/ner.12613

Matravers, M. (2000). *Justice and punishment: The rationale of coercion*. Oxford University Press on Demand.

Matravers, M. (2018). The Importance of Context in Thinking About Crime-Preventing Neurointerventions. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press.

Maturo, A. (2012). Medicalization: current concept and future directions in a bionic society [Article]. *Mens Sana Monogr*, *10*(1), 122-133. https://doi.org/10.4103/0973-1229.91587

Matusow, H., Dickman, S. L., Rich, J. D., Fong, C., Dumont, D. M., Hardin, C., Marlowe, D., & Rosenblum, A. (2013). Medication assisted treatment in US drug courts: results from a nationwide survey of availability, barriers and attitudes. *J Subst Abuse Treat*, *44*(5), 473-480. https://doi.org/10.1016/j.jsat.2012.10.004

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, *114*(48), 12714-12719.

Mayer, E. E. (1947). Prefrontal lobotomy and the courts. *J. Crim. L. & Criminology*, *38*, 576.

McAdams, D. P. (2013). *The redemptive self : stories Americans live by*. New York : Oxford University Press; Revised and expanded edition.

McAdams, D. P., & McLean, K. C. (2013). Narrative Identity. *Current Directions in Psychological Science*, *22*(3), 233-238. https://doi.org/10.1177/0963721413475622

McCoy, L. G., Brenna, C., Morgado, F., Chen, S., & Das, S. (2020). Neuroethics, Neuroscience, and the Project of Human Self-Understanding. *AJOB Neurosci*, *11*(3), 207-209. https://doi.org/10.1080/21507740.2020.1778127

McCoy, T. (2012). *Hard time: reforming the penitentiary in nineteenth-century Canada*. Athabasca University Press.

McKenna, D. J., Towers, G. H., & Abbott, F. (1984). Monoamine oxidase inhibitors in South American hallucinogenic plants: tryptamine and beta-carboline constituents of ayahuasca. *J Ethnopharmacol*, *10*(2), 195-223. https://doi.org/10.1016/0378-8741(84)90003-5

McKenna, M. S. (2000). Assessing Reasons - Responsive Compatibilism. *International Journal of Philosophical Studies*, *8*(1), 89-114. https://doi.org/10.1080/096725500341738

McKinney, C. D. (2022). *The Experience of Meditation and Healing in Practitioners of the Wim Hof Method* ProQuest Dissertations Publishing].

McKinnon, C. (2012). *Climate change and future justice: Precaution, compensation and triage*. Routledge.

McLaughlin, K. A., & Lambert, H. K. (2017). Child Trauma Exposure and Psychopathology: Mechanisms of Risk and Resilience. *Curr Opin Psychol*, *14*, 29-34. https://doi.org/10.1016/j.copsyc.2016.10.004

McMillan, J. (2014). The kindest cut? Surgical castration, sex offenders and coercive offers. *J Med Ethics*, *40*(9), 583-590. https://doi.org/10.1136/medethics-2012-101030

McNamara, R. K., & Carlson, S. E. (2006). Role of omega-3 fatty acids in brain development and function: potential implications for the pathogenesis and prevention of psychopathology. *Prostaglandins Leukot Essent Fatty Acids*, *75*(4-5), 329-349. https://doi.org/10.1016/j.plefa.2006.07.010

McNeill, F. (2012). Four forms of 'offender' rehabilitation: Towards an interdisciplinary perspective. *Legal and Criminological Psychology*, *17*(1), 18-36. https://doi.org/10.1111/j.2044-8333.2011.02039.x

McTernan, E. (2018a). Those Who Forget the Past. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice* (pp. 274-288). Oxford University Press.

McTernan, E. (2018b). Those Who Forget the Past: An Ethical Challenge from the History of Treating Deviance. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice* (1 ed.). Oxford University Press.

Mears, D. P. (2012). *American Criminal Justice Policy*. New York: Cambridge University Press. https://doi.org/10.1017/cbo9780511794858

Meffert, H., Gazzola, V., den Boer, J. A., Bartels, A. A., & Keysers, C. (2013). Reduced spontaneous but relatively normal deliberate vicarious representations in psychopathy. *Brain*, *136*(Pt 8), 2550-2562. https://doi.org/10.1093/brain/awt190

Mele, A. R. (1987). *Irrationality : an essay on akrasia, self-deception, and self-control*. Oxford University Press.

Mele, A. R. (1992). *Springs of Action: Understanding Intentional Behavior*. Cary: Oxford University Press, Incorporated.

Mele, A. R. (1995). *Autonomous agents: From self-control to autonomy*. Oxford University Press.

Mele, A. R. (2006). *Free Will and Luck*. New York: Oxford University Press. https://doi.org/10.1093/0195305043.001.0001

Menzel Jr, E. W. (1974). A group of young chimpanzees in a one-acre field. In *Behavior of nonhuman primates* (Vol. 5, pp. 83-153). Elsevier.

Meriney, S. D. (2019). *Synaptic transmission*. London, England: Academic Press.

Merkel, R., Boer, G., Fegert, J., Galert, T., Hartmann, D., Nuttin, B., & Rosahl, S. (2007). *Intervening in the Brain*

Changing Psyche and Society (1st ed.)

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, *30*(1), 63-81.

Merlin, M. D. (2003). Archaeological evidence for the tradition of psychoactive plant use in the old world. *Economic Botany*, *57*(3), 295-323.

Mesulam, M. M. (1998). From sensation to cognition. *Brain*, *121 (Pt 6)*(6), 1013-1052. https://doi.org/10.1093/brain/121.6.1013

Meyers, D. (2000). Intersectional identity and the authentic self: Opposites attract! In C. Mackenzie & N. Stoljar (Eds.), *Relational Autonomy : Feminist Perspectives on Autonomy, Agency, and the Social Self*. Oxford University Press, Incorporated. http://ebookcentral.proquest.com/lib/ucalgary-ebooks/detail.action?docID=430598

Michaelian, K. (2016). *Mental time travel: Episodic memory and our knowledge of the personal past*. MIT Press.

Michaelian, K., Klein, S. B., & Szpunar, K. K. (2016). *Seeing the Future*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190241537.001.0001

Mielau, J., Vogel, M., Gutwinski, S., & Mick, I. (2021). New Approaches in Drug Dependence: Opioids. *Curr Addict Rep*, *8*(2), 298-305. https://doi.org/10.1007/s40429-021-00373-9

Mikhail, J. (2009). Moral grammar and intuitive jurisprudence: A formal model of unconscious moral and legal knowledge. *Psychology of learning and motivation*, *50*, 27-100.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.

Mikula, G., Scherer, K. R., & Athenstaedt, U. (1998). The role of injustice in the elicitation of differential emotional reactions. *Personality and social psychology bulletin*, *24*(7), 769-783.

Mill, J. S., & Mill, J. S. ([1859] 1966). *On liberty*. Springer.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu Rev Neurosci*, *24*(1), 167-202. https://doi.org/10.1146/annurev.neuro.24.1.167

Milton, J. (1667). *Paradise Lost*. Samuel Simmons.

Milton, J. ([1667] 2014). *Paradise lost*. Minneapolis, MN : First Avenue Editions, a division of Lerner Publishing Group.

Milton, J., & Hughes, M. Y. (1957). *Complete poems and major prose*. Odyssey Press.

Minow, M. (1990). *Making all the difference: Inclusion, exclusion, and American law*. Cornell University Press.

Mitchell, J. P., Schirmer, J., Ames, D. L., & Gilbert, D. T. (2011). Medial prefrontal cortex predicts intertemporal choice. *J Cogn Neurosci*, *23*(4), 857-866. https://doi.org/10.1162/jocn.2010.21479

Mitchell, O., Wilson, D. B., Eggers, A., & MacKenzie, D. L. (2012). Assessing the effectiveness of drug courts on recidivism: A meta-analytic review of traditional and non-traditional drug courts. *Journal of Criminal Justice*, *40*(1), 60-71.

Moghadisian, M. H., & Eskin, N. A. M. (2012). *Functional foods and cardiovascular disease*. Boca Raton, Fla. : CRC Press.

Molero-Chamizo, A., Martin Riquel, R., Moriana, J. A., Nitsche, M. A., & Rivera-Urbina, G. N. (2019). Bilateral Prefrontal Cortex Anodal tDCS Effects on Self-reported Aggressiveness in Imprisoned Violent Offenders. *Neuroscience*, *397*, 31-40. https://doi.org/10.1016/j.neuroscience.2018.11.018

Moles, A. (2014). The Public Ecology of Freedom of Association. *Res Publica*, *20*(1), 85-103.

Moll, J., de Oliveira-Souza, R., & Eslinger, P. J. (2003). Morals and the human brain: a working model. *Neuroreport*, *14*(3), 299-305. https://doi.org/10.1097/00001756-200303030-00001

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

Moore, K. D. (1997). *Pardons: Justice, mercy, and the public interest*. Oxford University Press.

Moore, M. S. (1987). The moral worth of retribution. *Responsibility, character, and the emotions: New essays in moral psychology*, 179-219.

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review*, *5*(2), 119-124. https://doi.org/10.1177/1754073912468165

Morris, B. (2021). Health & Wellness: A Cold Comfort --- Wim Hof Puts Stress On Ice. *The Wall Street journal. Eastern edition*.

Morris, C. W. (1991). Punishment and Loss of Moral Standing 1. *Canadian Journal of Philosophy*, *21*(1), 53-77.

Morris, H. (1968). Persons and punishment. *The Monist*, 475-501.

Morris, N. (1973). Future of Imprisonment: Toward a Punitive Philosophy, The. *Mich. L. Rev.*, *72*, 1161.

Morse, S. (2011). Neuroscience and the Future of Personhood and Responsibility. In *Constitution 3.0: Freedom and technological change* (pp. 113-129). Brookings Institution Press.

Morse, S. J. (2000). Rationality and responsibility. *SOUTHERN CALIF LAW R*, *74*(1), 251-268.

Morse, S. J. (2004). Reason, results, and criminal responsibility. *U ILLINOIS LAW REV*, *2004*(2), 363-444.

Morse, S. J. (2005). Brain overclaim syndrome and criminal responsibility: A diagnostic note. *Ohio St. J. Crim. L.*, *3*, 397.

Morse, S. J., Roskies, A. L., & John, D. a. C. T. M. F. (2013). *A primer on criminal law and neuroscience : a contribution of the law and neuroscience project supported by the MacArthur Foundation*. New York : Oxford University Press.

Morsink, J. (1999). *The Universal Declaration of Human Rights: origins, drafting, and intent*. university of Pennsylvania Press.

Morsink, J. (2009). *Inherent human rights: Philosophical roots of the universal declaration*. University of Pennsylvania Press.

Moseley, R. (2020). Morality and the Brain It's going to take all the brainpower we have to figure it out. *NAT HIST*, *128*(3), 28-33.

Moskal, J. R., Burch, R., Burgdorf, J. S., Kroes, R. A., Stanton, P. K., Disterhoft, J. F., & Leander, J. D. (2014). GLYX-13, an NMDA receptor glycine site functional partial agonist enhances cognition and produces antidepressant effects without the psychotomimetic side effects of NMDA receptor antagonists. *Expert Opin Investig Drugs*, *23*(2), 243-254. https://doi.org/10.1517/13543784.2014.852536

Muller, J. L., Ganssbauer, S., Sommer, M., Dohnel, K., Weber, T., Schmidt-Wilcke, T., & Hajak, G. (2008). Gray matter changes in right superior temporal gyrus in criminal

psychopaths. Evidence from voxel-based morphometry. *Psychiatry Res, 163*(3), 213-222. https://doi.org/10.1016/j.pscychresns.2007.08.010

Murphy, E. R., & Greely, H. (2011). What Will Be the Limits of Neuroscience-based Mindreading in the Law? In (Vol. 1): Oxford University Press.

Murphy, J. G. (1973). Marxism and retribution. *Philosophy & Public Affairs*, 217-243.

Murphy, J. G. (1979). *Retribution, Justice, and Therapy*. Reidel.

Musen, K., & Zimbardo, P. (1992). *Quiet rage the Stanford prison experiment*. Insight Media [prod.].

Muzik, O., & Diwadkar, V. A. (2019). Hierarchical control systems for the regulation of physiological homeostasis and affect: Can their interactions modulate mood and anhedonia? *Neurosci Biobehav Rev*, *105*, 251-261. https://doi.org/10.1016/j.neubiorev.2019.08.015

Muzik, O., Reilly, K. T., & Diwadkar, V. A. (2018). "Brain over body"-A study on the willful regulation of autonomic function during cold exposure. *neuroimage*, *172*, 632-641. https://doi.org/10.1016/j.neuroimage.2018.01.067

Nadelhoffer, T., Goya-Tocchetto, D., Wright, J. C., & McGuire, Q. Folk Jurisprudence and Neurointervention: An Interdisciplinary Investigation. In *Neurointerventions and the Law* (pp. 191-230). Oxford University Press.

Nadelhoffer, T., Goya-Tocchetto, D., Wright, J. C., & McGuire, Q. (2020). Folk Jurisprudence and Neurointervention: An Interdisciplinary Investigation. In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity*. Oxford University Press.

Nadelmann, E., & LaSalle, L. (2017). Two steps forward, one step back: current harm reduction policy and politics in the United States. *Harm Reduct J*, *14*(1), 37. https://doi.org/10.1186/s12954-017-0157-y

Nagel, T. (1974). What is it like to be a bat. *Readings in philosophy of psychology*, *1*, 159-168.

Nagel, T. (2012). *Mind and cosmos why the materialist neo-Darwinian conception of nature is almost certainly false*. New York ; Oxford : Oxford University Press.

Nagin, D. S., Cullen, F. T., & Jonson, C. L. (2009). Imprisonment and reoffending. *Crime and Justice*, *38*(1), 115-200.

Nedelsky, J. (1993). Reconceiving rights as relationship. *Rev. Const. Stud.*, *1*, 1.

Neier, A. (2013). *The international human rights movement: a history* (Vol. 14). Princeton University Press.

Nelken, D. (2009). Comparative Criminal Justice. *European journal of criminology*, *6*(4), 291-311. https://doi.org/10.1177/1477370809104684

Nettle, D., Harper, Z., Kidson, A., Stone, R., Penton-Voak, I. S., & Bateson, M. (2013). The watching eyes effect in the Dictator Game: it's not how much you give, it's being seen to give something. *Evolution and human behavior*, *34*(1), 35-40. https://doi.org/10.1016/j.evolhumbehav.2012.08.004

Neuhouser, F., & Neuhouser, F. (2009). *Foundations of Hegel's social theory: Actualizing freedom*. Harvard University Press.

Newman, D. (2019). Interpreting Freedom of Thought in the Canadian Charter of Rights and Freedoms. *Supreme Court Law Review*, *91*.

Newman, D. (2021). Freedom of Thought in Canada: The History of a Forgetting and the Potential of a Remembering. *European Journal of Comparative Law and Governance*, *8*(2-3), 226-244.

Newman, G. E., & Smith, R. K. (2016). Kinds of Authenticity. *Philosophy Compass*, *11*(10), 609-618. https://doi.org/10.1111/phc3.12343

Newton, A., May, X., Eames, S., & Ahmad, M. (2019). Economic and social costs of reoffending. *Ministry of Justice*.

Nickel, J. W. (1987). *Making sense of human rights: Philosophical reflections on the universal declaration of human rights*. Univ of California Press.

Nickel, J. W. (2008). Rethinking indivisibility: Towards a theory of supporting relations between human rights. *Hum. Rts. Q.*, *30*, 984-1001.

Nickel, J. W. (2010). Indivisibility and linkage arguments: A reply to Gilabert. *Human Rights Quarterly*, *32*(2), 439-446.

Nickel, J. W. (2016). Can a right to health care be justified by linkage arguments? *Theoretical medicine and bioethics*, *37*(4), 293-306.

Nietzsche, F. ([1882] [1974]). *The Gay Science* (W. Kaufmann, Trans.). Vintage.

Nietzsche, F. W. ([1844-1900] 2006). *Thus spoke Zarathustra : a book for all and none*. Cambridge : Cambridge University Press, 2006. https://search.library.wisc.edu/catalog/9910022139702121

Northoff, G., & Lamme, V. (2020). Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight? *Neurosci Biobehav Rev*, *118*, 568-587. https://doi.org/10.1016/j.neubiorev.2020.07.019

Northoff, G., & Wagner, N.-F. (2018). Personal Identity and Brain Identity. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Nozick, R. (1974). *Anarchy, State and Utopia*. Blackwell.

Nummenmaa, L., Lukkarinen, L., Sun, L., Putkinen, V., Seppala, K., Karjalainen, T., Karlsson, H. K., Hudson, M., Venetjoki, N., Salomaa, M., Rautio, P., Hirvonen, J., Lauerma, H., & Tiihonen, J. (2021). Brain Basis of Psychopathy in Criminal Offenders and General Population. *CEREB CORTEX*, *31*(9), 4104-4114. https://doi.org/10.1093/cercor/bhab072

Nussbaum, M. C. (1992). Human functioning and social justice: In defense of Aristotelian essentialism. *Political theory*, *20*(2), 202-246.

Nussbaum, M. C. (2009). *Frontiers of justice: Disability, nationality, species membership*. Harvard University Press.

Olsaretti, S. (2003). *Desert and justice*. Clarendon.

Orwell, G. (1977 [1949]). *1984*. Harcourt.

Ostermeier, M., & Schmoll, M. (2022). Exploring methods for targeted activation of the sympathetic nervous system without exercise. *Current directions in biomedical engineering*, *8*(3), 21-24. https://doi.org/10.1515/cdbme-2022-2006

Ostovar, T. (2009). *Childhood adversities as antecedents of suicide completion* (Publication Number MR61622) [M.Sc., McGill University (Canada)]. ProQuest Dissertations & Theses Global. Ann Arbor.

Owen, A. M. (2013). Detecting consciousness: a unique role for neuroimaging. *Annu Rev Psychol*, *64*(1), 109-133. https://doi.org/10.1146/annurev-psych-113011-143729

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science*, *313*(5792), 1402. https://doi.org/10.1126/science.1130197

Packer, A. M., Roska, B., & Hausser, M. (2013). Targeting neurons and photons for optogenetics. *NAT NEUROSCI*, *16*(7), 805-815. https://doi.org/10.1038/nn.3427

Pais-Vieira, M., Lebedev, M., Kunicki, C., Wang, J., & Nicolelis, M. A. (2013). A brain-to-brain interface for real-time sharing of sensorimotor information. *Sci Rep*, *3*(1), 1319. https://doi.org/10.1038/srep01319

Palk, A. C. (2018). Mandatory Neurointerventions Could Enhance the Mental Integrity of Certain Criminal Offenders. *AJOB neuroscience*, *9*(3), 150-152. https://doi.org/10.1080/21507740.2018.1496174

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press.

Panksepp, J. (2012). *The archaeology of mind : neuro-evolutionary origins of human emotions*. New York : W. W. Norton & Co.; 1st ed.

Panksepp, J., & Watt, D. (2011). What is Basic about Basic Emotions? Lasting Lessons from Affective Neuroscience. *Emotion Review*, *3*(4), 387-396. https://doi.org/10.1177/1754073911410741

Pappadopulos, E., Woolston, S., Chait, A., Perkins, M., Connor, D. F., & Jensen, P. S. (2006). Pharmacotherapy of aggression in children and adolescents: efficacy and effect size. *Journal of the Canadian Academy of Child and Adolescent Psychiatry/Journal de l'Académie canadienne de psychiatrie de l'enfant et de l'adolescent*, *15*, 27-39.

Parastarfeizabadi, M., & Kouzani, A. Z. (2017). Advances in closed-loop deep brain stimulation devices. *J Neuroeng Rehabil*, *14*(1), 79. https://doi.org/10.1186/s12984-017-0295-1

Pardo, M. S. (2014). *Minds, brains, and law : the conceptual foundations of law and neuroscience*. New York : Oxford University Press.

Parfit, D. (1984). *Reasons and persons*. Clarendon Press. Publisher description http://www.loc.gov/catdir/enhancements/fy0639/83015139-d.html

Park, H. J., & Friston, K. (2013). Structural and functional brain networks: from connections to cognition. *Science*, *342*(6158), 1238411. https://doi.org/10.1126/science.1238411

Park, R. J., Singh, I., Pike, A. C., & Tan, J. O. (2017). Deep Brain Stimulation in Anorexia Nervosa: Hope for the Hopeless or Exploitation of the Vulnerable? The Oxford Neuroethics Gold Standard Framework [Clinical Trial]. *Front Psychiatry*, *8*(44), 44. https://doi.org/10.3389/fpsyt.2017.00044

Pascual, L., Gallardo-Pujol, D., & Rodrigues, P. (2013). How does morality work in the brain? A functional and structural perspective of moral behavior. *Frontiers in Integrative Neuroscience*, *7*, 65-65.

Patten, A. (1999). *Hegel's idea of freedom*. Oxford University Press.

Paul, S., Beucke, J. C., Kaufmann, C., Mersov, A., Heinzel, S., Kathmann, N., & Simon, D. (2019). Amygdala-prefrontal connectivity during appraisal of symptom-related stimuli in obsessive-compulsive disorder. *Psychol Med*, *49*(2), 278-286. https://doi.org/10.1017/S003329171800079X

Pearson, A. R., Dovidio, J. F., & Gaertner, S. L. (2009). The nature of contemporary prejudice: Insights from aversive racism. *Social and Personality Psychology Compass*, *3*(3), 314-338.

Pedroni, A., Eisenegger, C., Hartmann, M. N., Fischbacher, U., & Knoch, D. (2014). Dopaminergic stimulation increases selfish behavior in the absence of punishment threat. *Psychopharmacology (Berl)*, *231*(1), 135-141. https://doi.org/10.1007/s00213-013-3210-x

Pellegrini, R. J., Schauss, A. G., & Miller, M. E. (1981). Room color and aggression in a criminal detention holding cell: A test of the "tranquilizing pink" hypothesis. *Journal of Othomolecular Psychiatry*, *10*(3), 174-181.

Pereboom, D. (2009). *Living without Free Will*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511498824

Pereboom, D. (2014). *Free will, agency, and meaning in life* (1 ed.). Oxford University Press.

Pereboom, D. (2018). Incapacitation, Reintegration, and Limited General Deterrence. *Neuroethics*, *13*(1), 87-97. https://doi.org/10.1007/s12152-018-9382-7

Persson, I., & Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of applied philosophy*, *25*(3), 162-177.

Persson, I., & Savulescu, J. (2011a). The turn for ultimate harm: a reply to Fenton. *J Med Ethics*, *37*(7), 441-444. https://doi.org/10.1136/jme.2010.036962

Persson, I., & Savulescu, J. (2011b). Unfit for the Future? Human Nature, Scientific Progress, and the Need for Moral Enhancement. In *Enhancing Human Capacities* (pp. 486-500). https://doi.org/10.1002/9781444393552.ch35

Persson, I., & Savulescu, J. (2012). *Unfit for the future: the need for moral enhancement*. Oxford University Press.

Persson, I., & Savulescu, J. (2013). Getting moral enhancement right: the desirability of moral bioenhancement. *BIOETHICS*, *27*(3), 124-131. https://doi.org/10.1111/j.1467-8519.2011.01907.x

Persson, I., & Savulescu, J. (2015). The art of misunderstanding moral bioenhancement. *Camb Q Healthc Ethics*, *24*(1), 48-57. https://doi.org/10.1017/S0963180114000292

Petersen, T. S., & Kragh, K. (2017). Should violent offenders be forced to undergo neurotechnological treatment? A critical discussion of the 'freedom of thought' objection. *J Med Ethics*, *43*(1), 30-34. https://doi.org/10.1136/medethics-2016-103492

Petraskova Touskova, T., Bob, P., Bares, Z., Vanickova, Z., Nyvlt, D., & Raboch, J. (2022). A novel Wim Hof psychophysiological training program to reduce stress responses during an Antarctic expedition. *J Int Med Res*, *50*(4), 3000605221089883. https://doi.org/10.1177/03000605221089883

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *J Cogn Neurosci*, *12*(5), 729-738. https://doi.org/10.1162/089892900562552

Philips, J. (2014). On setting priorities among human rights. *Human rights review*, *15*(3), 239-257.

Pincoffs, E. (1977). Are questions of desert decidable? *Justice and punishment*, 75-88.

Piotrowska, P. J., Stride, C. B., Croft, S. E., & Rowe, R. (2015). Socioeconomic status and antisocial behaviour among children and adolescents: a systematic review and meta-analysis. *CLIN PSYCHOL REV*, *35*, 47-55. https://doi.org/10.1016/j.cpr.2014.11.003

Plunkett, D., & Shapiro, S. (2017). Law, Morality, and Everything Else: General Jurisprudence as a Branch of Metanormative Inquiry. *Ethics*, *128*(1), 37-68. https://doi.org/10.1086/692941

Poeppl, T. B., Donges, M. R., Mokros, A., Rupprecht, R., Fox, P. T., Laird, A. R., Bzdok, D., Langguth, B., & Eickhoff, S. B. (2019). A view behind the mask of sanity: meta-analysis of aberrant brain activity in psychopaths. *Mol Psychiatry*, *24*(3), 463-470. https://doi.org/10.1038/s41380-018-0122-5

Pogge, T. (2005). World poverty and human rights. *Ethics & international affairs*, *19*(1), 1-7.

Pont, J., Stover, H., & Wolff, H. (2012). Dual loyalty in prison health care. *Am J Public Health*, *102*(3), 475-480. https://doi.org/10.2105/AJPH.2011.300374

Pouw, W. T. J. L., van Gog, T., & Paas, F. (2014). An Embedded and Embodied Cognition Review of Instructional Manipulatives. *Educational psychology review*, *26*(1), 51-72. https://doi.org/10.1007/s10648-014-9255-5

Powell, N. L., & Derbyshire, S. W. G. (2018). Values, Empathy, and the Brain. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Pratt, J. (2007a). Scandinavian Exceptionalism in an Era of Penal Excess: Part I: The Nature and Roots of Scandinavian Exceptionalism. *British Journal of Criminology*, *48*(2), 119-137. https://doi.org/10.1093/bjc/azm072

Pratt, J. (2007b). Scandinavian Exceptionalism in an Era of Penal Excess: Part II: Does Scandinavian Exceptionalism Have a Future? *British Journal of Criminology*, *48*(3), 275-292. https://doi.org/10.1093/bjc/azm073

Primoratz, I. (1997). *Justifying legal punishment*. Prometheus Books.

Proix, T., Truccolo, W., Leguia, M. G., Tcheng, T. K., King-Stephens, D., Rao, V. R., & Baud, M. O. (2021). Forecasting seizure risk in adults with focal epilepsy: a development and validation study. *LANCET NEUROL*, *20*(2), 127-135. https://doi.org/10.1016/S1474-4422(20)30396-3

Pronin, E., Olivola, C. Y., & Kennedy, K. A. (2008). Doing unto future selves as you would do unto others: Psychological distance and decision making. *Personality and social psychology bulletin*, *34*(2), 224-236.

Pronin, E., & Ross, L. (2006). Temporal differences in trait self-ascription: when the self is seen as an other. *Journal of personality and social psychology*, *90*(2), 197-209.

Pugh, J. (2018). Coercion and the Neurocorrective Offer. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press. https://doi.org/DOI: 10.1093/oso/9780198758617.003.0005

Pugh, J. (2019). Moral Bio-enhancement, Freedom, Value and the Parity Principle. *Topoi (Dordr)*, *38*(1), 73-86. https://doi.org/10.1007/s11245-017-9482-8

Pugh, J. (2020). *Autonomy, rationality, and contemporary bioethics*. Oxford University Press.

Pujol, J., Batalla, I., Contreras-Rodriguez, O., Harrison, B. J., Pera, V., Hernandez-Ribas, R., Real, E., Bosa, L., Soriano-Mas, C., Deus, J., Lopez-Sola, M., Pifarre, J., Menchon, J. M., & Cardoner, N. (2012). Breakdown in the brain network subserving moral judgment in criminal psychopathy. *Soc Cogn Affect Neurosci*, *7*(8), 917-923. https://doi.org/10.1093/scan/nsr075

Qadir, H., Krimmel, S. R., Mu, C., Poulopoulos, A., Seminowicz, D. A., & Mathur, B. N. (2018). Structural Connectivity of the Anterior Cingulate Cortex, Claustrum, and the Anterior Insula of the Mouse. *Front Neuroanat*, *12*, 100. https://doi.org/10.3389/fnana.2018.00100

Quante, M. (2011). In defence of personal autonomy. *J Med Ethics*, *37*(10), 597-600. https://doi.org/10.1136/jme.2010.035717

Quinn, W. (1985). The right to threaten and the right to punish. *Philosophy & Public Affairs*, 327-373.

Quintavalla, A., & Heine, K. (2019). Priorities and human rights. *The International Journal of Human Rights*, *23*(4), 679-697.

Racine, E., Bar-Ilan, O., & Illes, J. (2006). Brain Imaging: A Decade of Coverage in the Print Media. *SCI COMMUN*, *28*(1), 122-142. https://doi.org/10.1177/1075547006291990

Racine, E., & Dubljevic, V. (2016). Porous or Contextualized Autonomy? Knowledge Can Empower Autonomous Moral Agents. *Am J Bioeth*, *16*(2), 48-50. https://doi.org/10.1080/15265161.2015.1120800

Racine, E., Nguyen, V., Saigle, V., & Dubljevic, V. (2017). Media Portrayal of a Landmark Neuroscience Experiment on Free Will. *Sci Eng Ethics*, *23*(4), 989-1007. https://doi.org/10.1007/s11948-016-9845-3

Raine, A. (2008). From Genes to Brain to Antisocial Behavior. *Current Directions in Psychological Science*, *17*(5), 323-328. https://doi.org/10.1111/j.1467-8721.2008.00599.x

Rakic, V. (2019). Genome Editing for Involuntary Moral Enhancement. *Camb Q Healthc Ethics*, *28*(1), 46-54. https://doi.org/10.1017/S0963180118000373

Raus, K., Focquaert, F., Schermer, M., Specker, J., & Sterckx, S. (2014). On Defining Moral Enhancement: A Clarificatory Taxonomy. *Neuroethics*, *7*(3), 263-273. https://doi.org/10.1007/s12152-014-9205-4

Rawls, J. (1955). Two concepts of rules. *The Philosophical Review*, *64*(1), 3-32.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Rawls, J. (2001). *Justice as fairness: A restatement* (E. Kelly, Ed.). Harvard University Press.

Rawls, J. (2005). *Political Liberalism*. Columbia University Press.

Rawls, J. (2009). *A theory of justice*. Harvard university press.

Raynor, P., & Robinson, G. (2005). *Rehabilitation, crime and justice*. Springer.

Raz, J. (1984). On the nature of rights. *Mind*, *93*(370), 194-214.

Raz, J. (1999). *Practical reason and norms*. Oxford University Press.

Raz, J., & Heuer, U. (2022). *The roots of normativity* (First edition. ed.). Oxford University Press.

Reghunath, A., & Ghasi, R. G. (2020). A journey through formation and malformations of the neo-cortex. *Childs Nerv Syst*, *36*(1), 27-38. https://doi.org/10.1007/s00381-019-04429-0

Regier, D. A., Farmer, M. E., Rae, D. S., Locke, B. Z., Keith, S. J., Judd, L. L., & Goodwin, F. K. (1990). Comorbidity of mental disorders with alcohol and other drug abuse. Results from the Epidemiologic Catchment Area (ECA) Study. *Jama*, *264*(19), 2511-2518. https://www.ncbi.nlm.nih.gov/pubmed/2232018

Rehm, J., Marmet, S., Anderson, P., Gual, A., Kraus, L., Nutt, D. J., Room, R., Samokhvalov, A. V., Scafato, E., Trapencieris, M., Wiers, R. W., & Gmel, G. (2013). Defining substance use disorders: do we really need more than heavy use? *Alcohol Alcohol*, *48*(6), 633-640. https://doi.org/10.1093/alcalc/agt127

Reiner, P. B. (2011). The Rise of Neuroessentialism. In (Vol. 1): Oxford University Press.

Reisch, L. A., & Sunstein, C. R. (2016). Do Europeans like nudges? *Judgment and Decision making*, *11*(4), 310-325.

Resnick, I., Newcombe, N. S., & Shipley, T. F. (2017). Dealing with Big Numbers: Representation and Understanding of Magnitudes Outside of Human Experience. *Cogn Sci*, *41*(4), 1020-1041. https://doi.org/10.1111/cogs.12388

Resnik, D. B. (2003). Is the precautionary principle unscientific? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *34*(2), 329-344.

Resnik, D. B. (2004). The precautionary principle and medical decision making. *J Med Philos*, *29*(3), 281-299. https://doi.org/10.1080/03605310490500509

Richell, R. A., Mitchell, D. G., Newman, C., Leonard, A., Baron-Cohen, S., & Blair, R. J. R. (2003). Theory of mind and psychopathy: can psychopathic individuals read the 'language of the eyes'? *Neuropsychologia*, *41*(5), 523-526.

Riva, P., Romero Lauro, L. J., DeWall, C. N., Chester, D. S., & Bushman, B. J. (2015). Reducing aggressive responses to social exclusion using transcranial direct current

stimulation. *Soc Cogn Affect Neurosci*, *10*(3), 352-356. https://doi.org/10.1093/scan/nsu053

Robb, D. (2022). Moral Responsibility and the Principle of Alternative Possibilities. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Roberts, J. V., & Ashworth, A. (2016). The Evolution of Sentencing Policy and Practice in England and Wales, 2003–2015. *Crime and Justice*, *45*(1), 307-358. https://doi.org/10.1086/685754

Robinson, P. H. (2019). *Justice, liability, and blame: Community views and the criminal law*. Routledge.

Rodriguez Arce, J. M., & Winkelman, M. J. (2021). Psychedelics, Sociality, and Human Evolution [Hypothesis and Theory]. *FRONT PSYCHOL*, *12*(4333), 729425. https://doi.org/10.3389/fpsyg.2021.729425

Roese, N. J., & Vohs, K. D. (2012). Hindsight Bias. *Perspect Psychol Sci*, *7*(5), 411-426. https://doi.org/10.1177/1745691612454303

Rolls, E. T. (2015). Limbic systems for emotion and for memory, but no single limbic system. *Cortex*, *62*, 119-157. https://doi.org/10.1016/j.cortex.2013.12.005

Rose, N. S. (2007). *The politics of life itself : biomedicine, power, and subjectivity in the twenty-first century*. Princeton : Princeton University Press.

Roskies, A. (2002). Neuroethics for the new millenium. *Neuron*, *35*(1), 21-23. https://doi.org/10.1016/s0896-6273(02)00763-8

Roskies, A. (2020a). Neuroethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/neuroethics/

Roskies, A. (2020b). Neuroethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*: Metaphysics Research Lab, Stanford University.

Roskies, A. L. (2022). Are Neuroimages Like Photographs of the Brain? *Philosophy of science*, *74*(5), 860-872. https://doi.org/10.1086/525627

Ross, D., Spurrett, D., Stephens, G. L., & Kincaid, H. (2007). *Distributed cognition and the will: Individual volition and social context*. MIT Press.

Rousseau, J.-J. ([1762] 1964). The social contract (1762). *Londres*.

Rumble, W. E. (1832 [1995]). *Austin: The province of jurisprudence determined*. Cambridge University Press.

Rundle, S. M., Cunningham, J. A., & Hendershot, C. S. (2021). Implications of addiction diagnosis and addiction beliefs for public stigma: A cross-national experimental study. *Drug Alcohol Rev*, *40*(5), 842-846. https://doi.org/10.1111/dar.13244

Rupert, R. D. (2009). *Cognitive systems and the extended mind*. Oxford University Press.

Russoniello, K., Vakharia, S. P., Netherland, J., Naidoo, T., Wheelock, H., Hurst, T., & Rouhani, S. (2023). Decriminalization of drug possession in Oregon: Analysis and early lessons. *Drug Science, Policy and Law*, *9*, 20503245231167407.

Rüther, M., & Heinrichs, J.-H. (2019). Human Enhancement: Deontological Arguments. *Zeitschrift für Ethik und Moralphilosophie*, *2*(1), 161-178. https://doi.org/10.1007/s42048-019-00036-5

Ryan, C. J. (2020). Is It Really Ethical to Prescribe Antiandrogens to Sex Offenders to Decrease Their Risk of Recidivism? In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity* (pp. 270-292). Oxford University Press.

Ryberg, J. (2012). Punishment, Pharmacological Treatment, and Early Release. *International Journal of Applied Philosophy*, *26*(2), 231-244. https://doi.org/10.5840/ijap201226217

Ryberg, J. (2018). Neuroscientific Treatment of Criminals and Penal Theory. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press.

Ryberg, J. (2020). *Neurointerventions, crime, and punishment : ethical considerations*. Oxford University Press.

Ryberg, J. (2021). Neurointerventions and crime prevention. An ethically inappropriate discussion? *Teoria e Critica della Regolazione Sociale/Theory and Criticism of Social Regulation*, *1*(22), 193-207.

Ryberg, J., & Petersen, T. S. (2011). Neurotechnological Behavioural Treatment of Criminal Offenders—A Comment on Bomann-Larsen. *Neuroethics*, *6*(1), 79-83. https://doi.org/10.1007/s12152-011-9146-0

Ryle, G. (1949). *The Concept of Mind*. Hutchinson.

Rzesnitzek, L., & Lang, S. (2016). A Material History of Electroshock Therapy : Electroshock Technology in Europe until 1945. *NTM, 24*(3), 251-277. https://doi.org/10.1007/s00048-016-0152-5 (Eine Geschichte der Elektroschocktherapie ‚von unten'. Elektroschocktechnologie in Europa bis 1945.)

Rzesnitzek, L., & Lang, S. (2017). 'Electroshock Therapy' in the Third Reich. *Med Hist, 61*(1), 66-88. https://doi.org/10.1017/mdh.2016.101

Sahakian, B., & LaBuzetta, J. N. (2013). *Bad Moves: How decision making goes wrong, and the ethics of smart drugs*. OUP Oxford.

Sale, K. (1995). *Rebels against the future : the Luddites and their war on the Industrial Revolution : lessons for the computer age*. Reading, Mass. : Addison-Wesley Pub. Co.

Salles, A., Evers, K., & Farisco, M. (2018). Neuroethics and Philosophy in Responsible Research and Innovation: The Case of the Human Brain Project. *Neuroethics, 12*(2), 201-211. https://doi.org/10.1007/s12152-018-9372-9

Sandel, M. (2002). What's wrong with enhancement. *President's Council on Bioethics, Washington, DC (www. bioethics. gov), 12*.

Sandel, M. J. (2007). *The case against perfection : ethics in the age of genetic engineering*. Cambridge, Mass. : Belknap Press of Harvard University Press.

Sandin, P. (2009). A new virtue-based understanding of the precautionary principle. *The ethics of protocells: Moral and social implications of creating life in the laboratory*, 89-104.

Sapolsky, R. M. (2004). The frontal cortex and the criminal justice system. *Philos Trans R Soc Lond B Biol Sci, 359*(1451), 1787-1796. https://doi.org/10.1098/rstb.2004.1547

Satel, S. L., & Lilienfeld, S. O. (2013). *Brainwashed : the seductive appeal of mindless neuroscience*. Basic Books.

Savage, J. B. (1977). Freedom and Necessity in Paradise Lost. *ELH, 44*(2), 286-311. https://doi.org/10.2307/2872669

Savulescu, J., & Persson, I. (2012). Moral enhancement, freedom and the god machine. *The Monist, 95*(3), 399-421.

Sawyer, W., & Wagner, P. (2019). Mass incarceration: The whole pie 2019. *Prison policy initiative*, *19*.

Scanlon, T. (2000). *What we owe to each other*. Belknap Press.

Scanlon, T. (2014). *Being realistic about reasons*. Oxford University Press.

Scanlon, T. (2020). *Rights, Balancing, and Proportionality*. file:///C:/Users/PhD/Downloads/SSRN-id3462529.pdf

Scarpazza, C., Ferracuti, S., Miolla, A., & Sartori, G. (2018). The charm of structural neuroimaging in insanity evaluations: guidelines to avoid misinterpretation of the findings. *Transl Psychiatry*, *8*(1), 227. https://doi.org/10.1038/s41398-018-0274-8

Schaefer, G. O., Kahane, G., & Savulescu, J. (2014). Autonomy and Enhancement. *Neuroethics*, *7*(2), 123-136. https://doi.org/10.1007/s12152-013-9189-5

Schauss, A. G. (1979). Tranquilizing effect of color reduces aggressive behavior and potential violence. *Journal of Orthomolecular Psychiatry*, *8*(4), 218-221.

Schechtman, M. (2012). Making the truth: Self-understanding, self-constitution, neuroscience, and narrative. *AJOB neuroscience*, *3*(4), 75-76.

Scheid, D. E. (1980). Note on defining 'punishment'. *Canadian Journal of Philosophy*, *10*(3), 453-462.

Schermer, M. (2015). Reducing, restoring or enhancing autonomy with neuromodulation techniques. In W. Glannon (Ed.), *Free will and the brain : neuroscientific, philosophical, and legal perspectives* (pp. 205-228). Cambridge, United Kingdom : Cambridge University Press.

Scheutz, M., & Malle, B. F. (2018). Moral Robots. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Schick, A. (2005). Neuro exceptionalism? *Am J Bioeth*, *5*(2), 36-38; discussion W33-34. https://doi.org/10.1080/15265160590960410

Schleim, S., & Quednow, B. B. (2018). How Realistic Are the Scientific Assumptions of the Neuroenhancement Debate? Assessing the Pharmacological Optimism and Neuroenhancement Prevalence Hypotheses. *FRONT PHARMACOL*, *9*, 3. https://doi.org/10.3389/fphar.2018.00003

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Pers Soc Psychol Bull*, *34*(8), 1096-1109. https://doi.org/10.1177/0146167208317771

Schoenthaler Stephen Amos Walter Do, S. (2009). The Effect of Randomized Vitamin-Mineral Supplementation on Violent and Non-violent Antisocial Behavior Among Incarcerated Juveniles. *Journal of Nutritional & Environmental Medicine*, *7*(4), 343-352. https://doi.org/10.1080/13590849762475

Schonau, A., Dasgupta, I., Brown, T., Versalovic, E., Klein, E., & Goering, S. (2021). Mapping the Dimensions of Agency. *AJOB Neurosci*, *12*(2-3), 172-186. https://doi.org/10.1080/21507740.2021.1896599

Schroeder, M. (2007). *Slaves of the Passions*. Oxford University Press.

Schultz, W. (2015). Neuroessentialism: Theoretical and Clinical Considerations. *Journal of Humanistic Psychology*, *58*(6), 607-639. https://doi.org/10.1177/0022167815617296

Scott, C. L., & Holmberg, T. (2003). Castration of sex offenders: prisoners' rights versus public safety. *J Am Acad Psychiatry Law*, *31*(4), 502-509. https://www.ncbi.nlm.nih.gov/pubmed/14974806

Seaman, J. A. (2008). Black boxes. *Emory law journal*, *58*(2), 427.

Seaman, J. A. (2018). Your Brain on Lies. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge university press.

Sebastian, P. M., & Sahakian, B. J. (2018). Modafinil and the Increasing Lifestyle Use of Smart Drugs by Healthy People. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Sedlmeier, P., Eberth, J., Schwarz, M., Zimmermann, D., Haarig, F., Jaeger, S., & Kunze, S. (2012). The psychological effects of meditation: a meta-analysis. *PSYCHOL BULL*, *138*(6), 1139-1171. https://doi.org/10.1037/a0028168

Sen, A. (2001). *Development as freedom*. Oxford Paperbacks.

Sententia, W. (2004). Neuroethical considerations: cognitive liberty and converging technologies for improving human cognition. *Ann N Y Acad Sci*, *1013*(1), 221-228. https://doi.org/10.1196/annals.1305.014

Sententia, W. (2013). Freedom by Design: Transhumanist Values and Cognitive Liberty. In (pp. 355-360).

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, *132*(3), 617-627.

Shammas, V. L. (2014). The pains of freedom: Assessing the ambiguity of Scandinavian penal exceptionalism on Norway's Prison Island. *Punishment & Society*, *16*(1), 104-123. https://doi.org/10.1177/1462474513504799

Shaw, E. (2012). Direct Brain Interventions and Responsibility Enhancement. *Criminal Law and Philosophy*, *8*(1), 1-20. https://doi.org/10.1007/s11572-012-9152-2

Shaw, E. (2018). Against the Mandatory Use of Neurointerventions in Criminal Sentencing. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice* (pp. 321-337). Oxford University Press.

Shaw, E., Pereboom, D., & Caruso, G. D. (2019). *Free will skepticism in law and society : challenging retributive justice*. Cambridge : Cambridge University Press.

Shen, Y.-Q., Zhou, H.-X., Chen, X., Castellanos, F. X., & Yan, C.-G. (2020). Meditation effect in changing functional integrations across large-scale brain networks: Preliminary evidence from a meta-analysis of seed-based functional connectivity. *Journal of Pacific Rim Psychology*, *14*, e10. https://doi.org/10.1017/prp.2020.1

Shewmon, D. A., Holmes, G. L., & Byrne, P. A. (1999). Consciousness in congenitally decorticate children: developmental vegetative state as self-fulfilling prophecy. *DEV MED CHILD NEUROL*, *41*(6), 364-374. https://doi.org/10.1017/s0012162299000821

Shoemaker, S. (1963). *Self-knowledge and self-identity*. Cornell University Press.

Shook, J. R. (2012). Neuroethics and the Possible Types of Moral Enhancement. *AJOB neuroscience*, *3*(4), 3-14. https://doi.org/10.1080/21507740.2012.712602

Shue, H. (2020). *Basic rights: Subsistence, affluence, and US foreign policy*. Princeton University Press.

Siegel, A. M., Barrett, M. S., & Bhati, M. T. (2017). Deep Brain Stimulation for Alzheimer's Disease: Ethical Challenges for Clinical Research. *J Alzheimers Dis*, *56*(2), 429-439. https://doi.org/10.3233/JAD-160356

Siegel, D. J. (2012). *The developing mind how relationships and the brain interact to shape who we are*. New York : Guilford Press; 2nd ed.

Sifferd, K. L. (2020). Chemical Castration as Punishment. In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity* (pp. 290-318). Oxford University Press.

Silver, M. G. (2003). Eugenics and compulsory sterilization laws: Providing redress for the victims of a shameful era in United States history. *Geo. Wash. L. Rev.*, *72*, 862-892.

Simmons, A. J. (1994). *The Lockean theory of rights*. Princeton University Press.

Simmons, A. J. (2010). Ideal and Nonideal Theory. *Philosophy & Public Affairs*, *38*(1), 5-36. https://doi.org/10.1111/j.1088-4963.2009.01172.x

Simmons, H. (2010). *Moral desert: A critique*. University Press of America.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, 229-243.

Singer, P. (2005). Ethics and Intuitions. *The Journal of Ethics*, *9*(3-4), 331-352. https://doi.org/10.1007/s10892-005-3508-y

Singer, P. (2011). *Practical ethics*. Cambridge university press.

Singer, R. G. (1979). *Just deserts : sentencing based on equality and desert*. Cambridge, Mass. : Ballinger.

Singer, W. (2019). A Naturalistic Approach to the Hard Problem of Consciousness. *Front Syst Neurosci*, *13*, 58. https://doi.org/10.3389/fnsys.2019.00058

Sinnott-Armstrong, W. E. (2008). *Moral psychology, Vol 1: The evolution of morality: Adaptations and innateness*. Boston Review.

Sjöstrand, M., & Juth, N. (2014). Authenticity and psychiatric disorder: does autonomy of personal preferences matter? *Medicine, Health Care and Philosophy*, *17*(1), 115-122.

Skyrms, B. (2010). *Signals : evolution, learning, & information*. Oxford ; New York : Oxford University Press.

Smart, J. J. C. (1961). Free-Will, Praise and Blame. *Mind*, *70*(279), 291-306.

Smilansky, S. (1996). Responsibility and desert: defending the connection. *Mind*, *105*(417), 157-163.

Smilansky, S. (2000). *Free will and illusion*. OUP Oxford.

Smith, J. E. (1997). The pre-eminence of autonomy in bioethics. In *Human Lives* (pp. 182-195). Springer.

Smith, P. S., & Ugelvik, T. (2017). *Scandinavian Penal History, Culture and Prison Practice: Embraced by the Welfare State?* Springer.

Smith, R. (2015). *Prison conditions: Overcrowding, disease, violence, and abuse*. Simon and Schuster.

Snodgrass, G. M., Blokland, A. A., Haviland, A., Nieuwbeerta, P., & Nagin, D. S. (2011). Does the time cause the crime? An examination of the relationship between time served and reoffending in the Netherlands. *Criminology*, *49*(4), 1149-1194.

Snow, N. E. (2010). *Virtue as social intelligence: An empirically grounded theory*. Routledge.

Sobhani, M., & Bechara, A. (2011). A somatic marker perspective of immoral and corrupt behavior. *Soc Neurosci*, *6*(5-6), 640-652. https://doi.org/10.1080/17470919.2011.605592

Spaniol, E. D., Smith, W. B., Thomas, D. A., & Clark, D. B. (2020). Addressing the opioid crisis: social and behavioral research contributions at the National Institutes of Health. *Transl Behav Med*, *10*(2), 482-485. https://doi.org/10.1093/tbm/ibz038

Sparrow, R. (2013). Better Living Through Chemistry? A Reply to Savulescu and Persson on 'Moral Enhancement'. *Journal of applied philosophy*, *31*(1), 23-32. https://doi.org/10.1111/japp.12038

Sparrow, R. (2014). Egalitarianism and moral bioenhancement. *Am J Bioeth*, *14*(4), 20-28. https://doi.org/10.1080/15265161.2014.889241

Sparrow, R. J. (2014). (Im) moral technology? Thought experiments and the future of'mind control'. In *The future of bioethics: International dialogues* (pp. 113-119). Oxford University Press.

Speak, D. (2002). Fanning the Flickers of Freedom. *American philosophical quarterly (Oxford)*, *39*(1), 91-105.

Speak, D. (2005). Semi-compatibilism and stalemate. *Philosophical Explorations*, *8*(2), 95-102. https://doi.org/10.1080/13869790500091391

Spece, R. G. (1972). Conditioning and other technologies used to treat rehabilitate demolish prisoners and mental patients. *S. Cal. L. Rev.*, *45*, 616.

Specker, J., Focquaert, F., Sterckx, S., & Schermer, M. H. N. (2017). Forensic practitioners' expectations and moral views regarding neurobiological interventions in offenders with mental disorders. *BioSocieties*, *13*(1), 304-321. https://doi.org/10.1057/s41292-017-0069-9

Spence, C. (2020). On the Ethics of Neuromarketing and Sensory Marketing. In J. T. Martineau & E. Racine (Eds.), *Organizational Neuroethics: Reflections on the Contributions of Neuroscience to Management Theories and Business Practice* (pp. 9-29). pringer International Publishing.

Spence, S. (2009). *The actor's brain*. Oxford University Press. https://doi.org/10.1093/med/9780198526667.001.0001

Sporns, O. (2010). *Networks of the Brain*. MIT press.

Sporns, O. (2016). *Networks of the Brain*. MIT press.

Sporns, O., Tononi, G., & Kötter, R. (2005). The human connectome: a structural description of the human brain. *PLoS computational biology*, *1*(4), 0245-0251.

Squire, L. R. (2009). Memory and brain systems: 1969-2009. *J NEUROSCI*, *29*(41), 12711-12716. https://doi.org/10.1523/JNEUROSCI.3575-09.2009

Squire, L. R. (2012). *Fundamental neuroscience*. Place of publication not identified Elsevier; Fourth edition.

Srinivasan, N. (2019). *Meditation*. Cambridge, Massachusetts : Academic Press; 1st edition.

Stanton, S. J., Sinnott-Armstrong, W., & Huettel, S. A. (2016). Neuromarketing: Ethical Implications of its Use and Potential Misuse. *Journal of Business Ethics*, *144*(4), 799-811. https://doi.org/10.1007/s10551-016-3059-0

Ṣṭaṭman, D. (1997). *Virtue ethics:[a critical reader]*. Edinburgh Univ. Press.

Stefano, F. (2021). Neurocorrection. On the use of neurodevices for criminals. *Teoria e Critica della Regolazione Sociale / Theory and Criticism of Social Regulation*, *1*(22). https://mimesisjournals.com/ojs/index.php/tcrs/article/view/1342

Steiner, H. (2006). Moral rights. In *The Oxford handbook of ethical theory* (pp. 459-479). Oxford University Press Oxford, United Kingdom.

Steinert, S., Bublitz, J.-C., Jox, R., & Friedrich, O. (2018). Doing Things with Thoughts: Brain-Computer Interfaces and Disembodied Agency. *Philosophy & Technology*, *32*(3), 457-482. https://doi.org/10.1007/s13347-018-0308-4

Stemplowska, Z. (2018). Should Coercive Neurointerventions Target the Victims of Wrongdoing? In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press.

Stender, J., Gosseries, O., Bruno, M. A., Charland-Verville, V., Vanhaudenhuyse, A., Demertzi, A., Chatelle, C., Thonnard, M., Thibaut, A., Heine, L., Soddu, A., Boly, M., Schnakers, C., Gjedde, A., & Laureys, S. (2014). Diagnostic precision of PET imaging and functional MRI in disorders of consciousness: a clinical validation study. *Lancet*, *384*(9942), 514-522. https://doi.org/10.1016/S0140-6736(14)60042-8

Stevens, A., Berto, D., Heckmann, W., Kerschl, V., Oeuvray, K., van Ooyen, M., Steffan, E., & Uchtenhagen, A. (2005). Quasi-compulsory treatment of drug dependent offenders: an international literature review. *Subst Use Misuse*, *40*(3), 269-283. https://doi.org/10.1081/ja-200049159

Stevens, C., Lauinger, B., & Neville, H. (2009). Differences in the neural mechanisms of selective attention in children from different socioeconomic backgrounds: an event-related brain potential study. *Dev Sci*, *12*(4), 634-646. https://doi.org/10.1111/j.1467-7687.2009.00807.x

Steward, H. (2009). Fairness, Agency and the Flicker of Freedom. *Nous*, *43*(1), 64-93. https://doi.org/10.1111/j.1468-0068.2008.01696.x

Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.

Steyn, J. (2004). Guantanamo Bay: the legal black hole. *International & Comparative Law Quarterly*, *53*(1), 1-15.

Stolzenberg, J. (2008). The pure 'I will'must be able to accompany all of my desires: The problem of a deduction of the categories of freedom in Kant's Critique of Practical Reason. *Recht und Frieden in der Philosophie Kants: Akten des X. Internationalen Kant-Kongresses*, *3*.

Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies*, *75*(1-2), 5-24. https://doi.org/10.1007/bf00989879

Strawson, G. (2010). *Freedom and Belief*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199247493.001.0001

Strawson, P. F. (1974). *Freedom and resentment : and other essays*. Methuen.

Striedter, G. F. (2005). *Principles of brain evolution*. Sinauer associates.

Strike, C., & Watson, T. M. (2019). Losing the uphill battle? Emergent harm reduction interventions and barriers during the opioid overdose crisis in Canada. *Int J Drug Policy*, *71*, 178-182. https://doi.org/10.1016/j.drugpo.2019.02.005

Sturge, G., & Tunnicliffe, R. (2021). *UK Prison Population Statistics*. United Kingdom Parliament - House of Commons Library Retrieved from https://researchbriefings.files.parliament.uk/documents/SN04334/SN04334.pdf

Sumner, W. (1987). The moral foundation of rights. In: Oxford: Oxford University Press.

Sunstein, C. R. (2017). *Why Nudge?* Yale University Press. https://doi.org/10.12987/9780300206920

Sunstein, C. R., & Thaler, R. H. (2003). Libertarian Paternalism Is Not an Oxymoron. *The University of Chicago law review*, *70*(4), 1159. https://doi.org/10.2307/1600573

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Svoboda, E., McKinnon, M. C., & Levine, B. (2006). The functional neuroanatomy of autobiographical memory: a meta-analysis. *Neuropsychologia*, *44*(12), 2189-2208. https://doi.org/10.1016/j.neuropsychologia.2006.05.023

Swanton, C. (2003). *Virtue ethics: A pluralistic view*. Clarendon Press.

Szentagothai, J. (1993). Self-organization: the basic principle of neural functions. *Theor Med*, *14*(2), 101-116. https://doi.org/10.1007/BF00997270

Tadros, V. (2011). *The ends of harm: The moral foundations of criminal law*. OUP Oxford.

Talbott, W. J. (2010). *Human rights and human well-being*. Oxford University Press.

Tallis, R. (2016). *Aping mankind : neuromania, Darwinitis and the misrepresentation of humanity*. London ; New York : Routledge; 1 ed.

Tamburrini, C., & Ta¨nnsjo€, T. r. (2011). Enhanced Bodies. In *Enhancing Human Capacities* (pp. 274-290). https://doi.org/10.1002/9781444393552.ch20

Tancredi, L. (2005). *Hardwired behavior: What neuroscience reveals about morality*. Cambridge University Press.

Tang, Y. Y., Holzel, B. K., & Posner, M. I. (2015). The neuroscience of mindfulness meditation. *NAT REV NEUROSCI*, *16*(4), 213-225. https://doi.org/10.1038/nrn3916

Tasioulas, J. (2015). On the foundations of human rights. *Penultimate version of chapter to appear in Cruft, Liao, Renzo (eds), Philosophical Foundations of Human Rights (OUP, 2014)*.

Tassy, S., Oullier, O., Duclos, Y., Coulon, O., Mancini, J., Deruelle, C., Attarian, S., Felician, O., & Wicker, B. (2012). Disrupting the right prefrontal cortex alters moral judgement. *Soc Cogn Affect Neurosci*, *7*(3), 282-288. https://doi.org/10.1093/scan/nsr008

Tavris, C., & Aronson, E. (2007). *Mistakes were made (but not by me): Why we justify foolish beliefs, bad decisions, and hurtful acts*. Harcourt.

Taylor, C. (1973). Interpretation and the sciences of man. In *Explorations in Phenomenology* (pp. 47-101). Springer.

Taylor, C. (2015). *Hegel and modern society*. Cambridge University Press.

Taylor, R. (1966). *Action and purpose*. Prentice-Hall.

Tegmark, M. (2000). Importance of quantum decoherence in brain processes. *Physical review E*, *61*(4), 4194-4206.

Teicher, M. H., Anderson, C. M., & Polcari, A. (2012). Childhood maltreatment is associated with reduced volume in the hippocampal subfields CA3, dentate gyrus,

and subiculum. *Proceedings of the National Academy of Sciences*, *109*(9), E563-E572.

Ten, C. L. (1987). *Crime, Guilt, & Punishment*. Oxford University Press Oxford.

Terbeck, S., Kahane, G., McTavish, S., Savulescu, J., Levy, N., Hewstone, M., & Cowen, P. (2014). Corrigendum to 'Beta adrenergic blockade reduces utilitarian judgement' [Biol. Psychol. 92 (2013) 323–328]. *Biological Psychology*, *103*, 370. https://doi.org/10.1016/j.biopsycho.2014.11.007

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. *J Pers Soc Psychol*, *78*(5), 853-870. https://doi.org/10.1037//0022-3514.78.5.853

Thacher, D. (2006). The normative case study. *American journal of sociology*, *111*(6), 1631-1676.

Thair, H., Holloway, A. L., Newport, R., & Smith, A. D. (2017). Transcranial Direct Current Stimulation (tDCS): A Beginner's Guide for Design and Implementation. *Frontiers in neuroscience*, *11*, 641. https://doi.org/10.3389/fnins.2017.00641

Thaler, R. H. (2008). *Nudge : improving decisions about health, wealth, and happiness*. New Haven : Yale University Press.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge : improving decisions about health, wealth, and happiness*. Yale University Press.

Thayer, J. F., Åhs, F., Fredrikson, M., Sollers III, J. J., & Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, *36*(2), 747-756.

Thayer, J. F., & Lane, R. D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. *J Affect Disord*, *61*(3), 201-216. https://doi.org/10.1016/s0165-0327(00)00338-4

Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.

Thompson, J. (1985). The Trolley Problem. *Yale Law Journal*, *94*(6), 1395-1415.

Tiihonen, J., Rossi, R., Laakso, M. P., Hodgins, S., Testa, C., Perez, J., Repo-Tiihonen, E., Vaurio, O., Soininen, H., Aronen, H. J., Kononen, M., Thompson, P. M., & Frisoni, G. B. (2008). Brain anatomy of persistent violent offenders: more rather than less. *Psychiatry Res*, *163*(3), 201-212. https://doi.org/10.1016/j.pscychresns.2007.08.012

Tobey, D. L. (2003). What's really wrong with genetic enhancement: a second look at our posthuman future. *Yale JL & Tech.*, *6*, 54.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *NAT REV NEUROSCI*, *17*(7), 450-461. https://doi.org/10.1038/nrn.2016.44

Tononi, G., & Sporns, O. (2003). Measuring information integration. *BMC Neurosci*, *4*, 31. https://doi.org/10.1186/1471-2202-4-31

Tonry, M. (2008). Learning from the limitations of deterrence research. *Crime and Justice*, *37*(1), 279-311.

Tonry, M. (2011a). Less imprisonment is no doubt a good thing: More policing is not. *Criminology & Pub. Pol'y*, *10*, 137.

Tonry, M. (2011b). *Retributivism has a past: has it a future?* Oxford University Press. https://doi.org/DOI: 10.1093/acprof:oso/9780199798278.001.0001

Tonry, M. (2016). Equality and human dignity: The missing ingredients in American sentencing. *Crime and Justice*, *45*(1), 459-496.

Tonry, M. (2017). Making American Sentencing Just, Humane, and Effective. *Crime and Justice*, *46*(1), 441-504. https://doi.org/10.1086/688456

Tovino, S. A. (2007). Functional neuroimaging information: a case for neuro exceptionalism? *Florida State University law review*, *34*(2), 415.

Trappe, H. J. (2010). The effects of music on the cardiovascular system and cardiovascular health. *Heart*, *96*(23), 1868-1871. https://doi.org/10.1136/hrt.2010.209858

Tse, P. (2013). *The neural basis of free will criterial causation*. Cambridge, MA : The MIT Press.

Tse, P. (2018). Two Types of Libertarian Free Will Are Realized in the Human Brain. In *Neuroexistentialism*. Oxford University Press. https://doi.org/10.1093/oso/9780190460723.003.0010

Tuck, R. (1979). *Natural rights theories: their origin and development*. Cambridge University Press.

Tufekci, Z. (2014). Engineering the public: Big data, surveillance and computational politics. *First Monday*.

Turner, D. C., Robbins, T. W., Clark, L., Aron, A. R., Dowson, J., & Sahakian, B. J. (2003). Cognitive enhancing effects of modafinil in healthy volunteers. *Psychopharmacology (Berl)*, *165*(3), 260-269. https://doi.org/10.1007/s00213-002-1250-8

Turner, D. C., & Sahakian, B. J. (2006). Neuroethics of Cognitive Enhancement. *BioSocieties*, *1*(1), 113-123. https://doi.org/10.1017/s1745855205040044

Turner, R. (2016). The Impact of Drug Treatment and Testing Orders in West Yorkshire: Six-Month Outcomes. *Probation Journal*, *51*(2), 116-132. https://doi.org/10.1177/0264550504044170

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207-232.

Ugelvik, T., & Dullum, J. (2012). *Penal exceptionalism? : nordic prison policy and practice*. Abingdon, Oxon ; New. York : Routledge.

Uhlmann, E., & Cohen, G. L. (2005). Constructed criteria: redefining merit to justify discrimination. *Psychol Sci*, *16*(6), 474-480. https://doi.org/10.1111/j.0956-7976.2005.01559.x

Ulman, Y. I., Cakar, T., & Yildiz, G. (2015). Ethical Issues in Neuromarketing: "I Consume, Therefore I am!". *Sci Eng Ethics*, *21*(5), 1271-1284. https://doi.org/10.1007/s11948-014-9581-5

Uludag, K., Ugurbil, K., & Berliner, L. (2015). *fMRI: from nuclear spins to brain functions* (Vol. 30). Springer.

Unger, P. K. (1996). *Living high and letting die: Our illusion of innocence*. Oxford University Press, USA.

Urban, J. H., & Rosenkranz, J. A. (2020). *Handbook of amygdala structure and function*. London : Academic Press.

Vallentyne, P. (2018a). Neurointervention, self-ownership, and enforcement rights. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press.

Vallentyne, P. (2018b). Neurointerventions, Self-Ownership, and Enforcement Rights. In D. Birks & T. Douglas (Eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice* (pp. 124-139). Oxford University Press.

Vallentyne, P. (2018c). Neurointerventions: Punishment, Mental Integrity, and Intentions. *AJOB neuroscience*, *9*(3), 131-132. https://doi.org/10.1080/21507740.2018.1496185

Vanderzyl, K. A. (1994). Castration as an alternative to incarceration: An impotent approach to the punishment of sex offenders. *N. Ill. UL Rev.*, *15*, 107.

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind, revised edition: Cognitive science and human experience*. MIT press.

Varga, S. (2015). Habermas' "Species Ethics", and the Limits of "Formal Anthropology". *Critical Horizons*, *12*(1), 71-89. https://doi.org/10.1558/crit.v12i1.71

Vargas, M. (2013). *Building better beings: A theory of moral responsibility*. OUP Oxford.

Vasiljevic, M., & Viki, G. T. (2013). Dehumanization, Moral Disengagement, and Public Attitudes to Crime and Punishment. In J. P. Leyens, P. G. Bain, & J. Vaes (Eds.), *Humanness and Dehumanization* (pp. 137-154). Psychology Press. https://doi.org/10.4324/9780203110539-14

Vaughan, B. (2006). The Internal Narrative of Desistance. *British Journal of Criminology*, *47*(3), 390-404. https://doi.org/10.1093/bjc/azl083

Veselis, R. A. (2017). The Memory Labyrinth: Systems, Processes, and Boundaries. In *Total Intravenous Anesthesia and Target Controlled Infusions* (pp. 31-62). Springer.

Viale, R. (2022). *Nudging* (1st ed.). MIT Press.

Vidal, F. (2009). Brainhood, anthropological figure of modernity. *Hist Human Sci*, *22*(1), 5-36. https://doi.org/10.1177/0952695108099133

Vidal, F. (2018). What makes neuroethics possible? *History of the human sciences*, *32*(2), 32-58. https://doi.org/10.1177/0952695118800410

Vihvelin, K. (2022). Arguments for Incompatibilism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Viki, G. T., Fullerton, I., Raggett, H., Tait, F., & Wiltshire, S. (2012). The role of dehumanization in attitudes toward the social exclusion and rehabilitation of sex offenders. *Journal of Applied Social Psychology*, *42*(10), 2349-2367.

Vincent, N. A. (2014). Restoring responsibility: Promoting justice, therapy and reform through direct brain interventions. *Criminal Law and Philosophy*, *8*(1), 21-42.

Vincent, N. A., Nadelhoffer, T., & McCay, A. (2020a). Law Viewed Through the Lens of Neurointerventions. In N. A. Vincent, T. Nadelhoffer, & A. McCay (Eds.), *Neurointerventions and the Law: Regulating Human Mental Capacity*. Oxford University Press.

Vincent, N. A., Nadelhoffer, T., & McCay, A. (2020). *Neurointerventions and the Law*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190651145.001.0001

Vincent, N. A., Nadelhoffer, T., & McCay, A. (2020b). *Neurointerventions and the Law: Regulating Human Mental Capacity*. Oxford University Press.

Violante, I. R., Alania, K., Cassarà, A. M., Neufeld, E., Acerbo, E., Carron, R., Williamson, A., Kurtin, D. L., Rhodes, E., Hampshire, A., Kuster, N., Boyden, E. S., Pascual-Leone, A., & Grossman, N. (2023). Non-invasive temporal interference electrical stimulation of the human hippocampus. *Nature Neuroscience*, *26*(11), 1994-2004. https://doi.org/10.1038/s41593-023-01456-8

Vogeley, K., Kurthen, M., Falkai, P., & Maier, W. (1999). Essential functions of the human self model are implemented in the prefrontal cortex. *Conscious Cogn*, *8*(3), 343-363. https://doi.org/10.1006/ccog.1999.0394

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: encouraging a belief in determinism increases cheating. *Psychol Sci*, *19*(1), 49-54. https://doi.org/10.1111/j.1467-9280.2008.02045.x

Von Hirsch, A. (1992). Proportionality in the Philosophy of Punishment. *Crime and Justice*, *16*, 55-98.

Von Hirsch, A. (1996). Censure and sanctions.

Von Hirsch, A. (2017). *Deserved criminal sentences*. Bloomsbury Publishing.

von Hirsch, A., & Ashworth, A. (2005). *Proportionate Sentencing*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199272600.001.0001

Wagner, D. D., Haxby, J. V., & Heatherton, T. F. (2012). The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdiscip Rev Cogn Sci*, *3*(4), 451-470. https://doi.org/10.1002/wcs.1183

Wagner, P., & Rabuy, B. (2017). Following the money of mass incarceration. *Prison policy initiative, 25*.

Walker, M. (2009). Enhancing genetic virtue. *Politics Life Sci*, *28*(2), 27-47. https://doi.org/10.2990/28_2_27

Walker, M. J., & Mackenzie, C. (2020). Neurotechnologies, Relational Autonomy, and Authenticity. *IJFAB: International Journal of Feminist Approaches to Bioethics*, *13*(1), 98-119. https://doi.org/10.3138/ijfab.13.1.06

Walker, N. (1991). *Why punish?* Oxford University Press Oxford.

Wallace, B. A. (1999). The Buddhist tradition of Samatha: Methods for refining and examining consciousness. *Journal of Consciousness studies*, *6*(3), 175-187.

Wallace, B. A., & Shapiro, S. L. (2006). Mental balance and well-being: building bridges between Buddhism and Western psychology. *AM PSYCHOL*, *61*(7), 690-701. https://doi.org/10.1037/0003-066X.61.7.690

Waller, B. N. (2015). *The stubborn system of moral responsibility*. MIT Press.

Walsh, C. (2010). Drugs and human rights: private palliatives, sacramental freedoms and cognitive liberty. *The International Journal of Human Rights*, *14*(3), 425-441. https://doi.org/10.1080/13642980802704270

Wang, D. J., Rao, H., Korczykowski, M., Wintering, N., Pluta, J., Khalsa, D. S., & Newberg, A. B. (2011). Cerebral blood flow changes associated with different meditation practices and perceived depth of meditation. *Psychiatry Res*, *191*(1), 60-67. https://doi.org/10.1016/j.pscychresns.2010.09.011

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, *12*(3), 129-140.

Watters, C. (2021). The Epidemic Confined Within the Pandemic: The American Prison System Must View Opioid Use Disorder as a Health Crisis During the COVID-19 Pandemic and Beyond. *Law & psychology review*, *45*, 243.

Wedgwood, R. (2007). Thinking About What Ought To Be. In *The Nature of Normativity*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199251315.003.0002

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. guilford Press.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *J Cogn Neurosci*, *20*(3), 470-477. https://doi.org/10.1162/jocn.2008.20040

Weizhi, D. (2010). Prisoners vs. the institution: Resistance in The Shawshank Redemption. *Folio*, *9*, 36-43.

Weller, J. A., Levin, I. P., Shiv, B., & Bechara, A. (2007). Neural correlates of adaptive decision making for risky gains and losses. *Psychol Sci*, *18*(11), 958-964. https://doi.org/10.1111/j.1467-9280.2007.02009.x

Wellman, C. (1987). A theory of rights: Persons under laws, institutions, and morals. *Ethics*, *97*(2), 474-476.

Wellman, C. H. (2009). Rights and state punishment. *The Journal of philosophy, 106*(8), 419-439.

Wellman, C. H. (2012). The rights forfeiture theory of punishment. *Ethics*, *122*(2), 371-393.

Wellman, C. H. (2017). *Rights forfeiture and punishment*. Oxford University Press.

Wellman, C. H. (2020). Clarifying Forfeiture Theory in Response to Dempsey and Lang. *Criminal Law and Philosophy*, 1-8.

Wellman, H. M. (2011). Developing a theory of mind.

Wexler, A., & Reiner, P. B. (2018). Home Use of tDCS. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Wheatley, T., & Decety, J. (2015). *The moral brain: A multidisciplinary perspective*. MIT Press.

White, M. D. (2011). *Retributivism: Essays on theory and policy*. Oxford University Press.

White, M. D., Saunders, J., Fisher, C., & Mellow, J. (2012). Exploring inmate reentry in a local jail setting: Implications for outreach, service use, and recidivism. *Crime & Delinquency*, *58*(1), 124-146.

Whitehead, A. N. (1929/1978). *Process and Reality*. Free Press.

Whitehead, R., & Chandler, J. A. (2018). Biocriminal Justice: Exploring Public Attitudes to Criminal Rehabilitation Using Biomedical Treatments. *Neuroethics*, *13*(1), 55-71. https://doi.org/10.1007/s12152-018-9370-y

Whitehouse, H. (2004). *Modes of religiosity: A cognitive theory of religious transmission*. Rowman Altamira.

Widerker, D., & McKenna, M. (2003). *Moral responsibility and alternative possibilities : essays on the importance of alternative possibilities*. Ashgate.

Widom, C. S. (1978). An empirical classification of female offenders. *Crim. Just. & Behavior*, *5*, 35.

Wilkinson, T. M. (2012). Nudging and Manipulation. *Political studies*, *61*(2), 341-355. https://doi.org/10.1111/j.1467-9248.2012.00974.x

Wilson, T. D. (2002). *Strangers to ourselves : discovering the adaptive unconscious*. Cambridge : Belknap Press of Harvard University Press.

Winkielman, P., Berridge, K. C., & Wilbarger, J. L. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Pers Soc Psychol Bull*, *31*(1), 121-135. https://doi.org/10.1177/0146167204271309

Wiseman, H. (2016). *The myth of the moral brain: The limits of moral enhancement*. MIT Press.

Witt, K. (2017). Identity change and informed consent. *J Med Ethics*, *43*(6), 384-390. https://doi.org/10.1136/medethics-2016-103684

Wolf, S. (1990). *Freedom within Reason*. Oxford University Press.

Wolff, J. (2011). *Ethics and public policy : a philosophical inquiry* (2nd ed.). Routledge.

Wolff, N., Fabrikant, N., & Belenko, S. (2011). Mental health courts and their selection processes: Modeling variation for consistency. *Law and Human Behavior*, *35*(5), 402-412.

Wolpaw, J., & Wolpaw, E. W. (2012). *Brain–Computer InterfacesPrinciples and Practice*. Cary: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195388855.001.0001

Wolpe, P. R. (2009). Is my mind mine? Neuroethics and brain imaging.

Wolpe, P. R. (2018). Neuroprivacy and Cognitive Liberty. In L. S. Johnson & K. S. Rommelfanger (Eds.), *The Routledge Handbook of Neuroethics*. Routledge.

Wood, D. (2002). Retribution, crime reduction and the justification of punishment. *Oxford Journal of Legal Studies*, *22*(2), 301-321.

Wu, Y.-C., Liao, Y.-S., Yeh, W.-H., Liang, S.-F., & Shaw, F.-Z. (2021). Directions of Deep Brain Stimulation for Epilepsy and Parkinson's Disease. *Frontiers in neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.680938

Wudarczyk, O. A., Earp, B. D., Guastella, A., & Savulescu, J. (2013). Could intranasal oxytocin be used to enhance relationships? Research imperatives, clinical policy, and ethical considerations. *Curr Opin Psychiatry*, *26*(5), 474-484. https://doi.org/10.1097/YCO.0b013e3283642e10

Yakobi, O., Smilek, D., & Danckert, J. (2021). The Effects of Mindfulness Meditation on Attention, Executive Control and Working Memory in Healthy Adults: A Meta-analysis of Randomized Controlled Trials. *Cognitive Therapy and Research*, *45*(4), 543-560. https://doi.org/10.1007/s10608-020-10177-2

Yang, Y., Raine, A., Narr, K. L., Colletti, P., & Toga, A. W. (2009). Localization of deformations within the amygdala in individuals with psychopathy. *Arch Gen Psychiatry*, *66*(9), 986-994. https://doi.org/10.1001/archgenpsychiatry.2009.110

Yehuda, S., Rabinovitz, S., & Mostofsky, D. I. (2005). Essential fatty acids and the brain: from infancy to aging. *Neurobiol Aging*, *26 Suppl 1*(1, Supplement), 98-102. https://doi.org/10.1016/j.neurobiolaging.2005.09.013

Yizhar, O., Fenno, L. E., Davidson, T. J., Mogri, M., & Deisseroth, K. (2011). Optogenetics in neural systems. *Neuron*, *71*(1), 9-34. https://doi.org/10.1016/j.neuron.2011.06.004

Yoo, S. S., Kim, H., Filandrianos, E., Taghados, S. J., & Park, S. (2013). Non-invasive brain-to-brain interface (BBI): establishing functional links between two brains. *PLOS ONE*, *8*(4), e60410. https://doi.org/10.1371/journal.pone.0060410

Young, G. (2019). Objections to the God Machine Thought Experiment and What they Reveal about the Intelligibility of Moral Intervention by Technological Means. *Philosophia*, *48*(2), 831-846. https://doi.org/10.1007/s11406-019-00095-3

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proc Natl Acad Sci U S A*, *107*(15), 6753-6758. https://doi.org/10.1073/pnas.0914826107

Young, L., & Dungan, J. (2012). Where in the brain is morality? Everywhere and maybe nowhere. *Social neuroscience*, *7*(1), 1-10.

Yowell, P. (2007). Critical Examination of Dworkin's Theory of Rights. *Am. J. Juris.*, *52*, 93-137.

Yue, C., Long, Y., Ni, C., Peng, C., & Yue, T. (2021). Valence of Temporal Self-Appraisals: A Comparison Between First-Person Perspective and Third-Person Perspective. *FRONT PSYCHOL*, *12*, 778532. https://doi.org/10.3389/fpsyg.2021.778532

Yuste, R., Goering, S., Arcas, B. A. Y., Bi, G., Carmena, J. M., Carter, A., Fins, J. J., Friesen, P., Gallant, J., Huggins, J. E., Illes, J., Kellmeyer, P., Klein, E., Marblestone, A., Mitchell, C., Parens, E., Pham, M., Rubel, A., Sadato, N., . . . Wolpaw, J. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*, *551*(7679), 159-163. https://doi.org/10.1038/551159a

Zaalberg, A., Nijman, H., Bulten, E., Stroosma, L., & van der Staak, C. (2010). Effects of nutritional supplements on aggression, rule-breaking, and psychopathology among young adult prisoners. *Aggress Behav*, *36*(2), 117-126. https://doi.org/10.1002/ab.20335

Zawadzki, P., & Adamczyk, A. K. (2021). Personality and Authenticity in Light of the Memory-Modifying Potential of Optogenetics. *AJOB Neurosci*, *12*(1), 3-21. https://doi.org/10.1080/21507740.2020.1866097

Zimbardo, P. (1983). To control a mind

Stanford Magazine, *11*, 59–64.

Zimbardo, P. (2007). *The Lucifer effect : understanding how good people turn evil*. New York : Random House; 1st ed.

Zimbardo, P. G., & Haney, C. (2020). Continuing to acknowledge the power of dehumanizing environments: Comment on Haslam et al. (2019) and Le Texier (2019). *AM PSYCHOL*, *75*(3), 400-402. https://doi.org/10.1037/amp0000593

Zimmerman, M. J. (2011). *The immorality of punishment*. Broadview Press.

Zuk, P., Torgerson, L., Sierra-Mercado, D., & Lazaro-Munoz, G. (2018). Neuroethics of Neuromodulation: An Update. *Curr Opin Biomed Eng*, *8*, 45-50. https://doi.org/10.1016/j.cobme.2018.10.003

Notes

1 In this dissertation, bibliographic information will be formatted following the guidelines of the American Psychological Association (APA) 6th edition. References to jurisprudence and statutory instruments will adhere to the standards outlined in *The Canadian Guide to Uniform Legal Citation*, 9th edition (Toronto: Carswell, 2018), which is the accepted legal citation reference system in Canada.

2 The question of how much control humans have over their biological nature is a complex issue which cannot be addressed here but traces themes that follow. I thank Walter Glannon for identifying this point.

3 The author extends gratitude to the committee for their insightful feedback, which recommended the formalization of PEAs, and potential formalizations, resulting in substantial changes to this introductory chapter. This suggestion has substantially enhanced the clarity and quality of the work, allowing for a more nuanced exploration of the ethical considerations surrounding compulsory neurointerventions within the criminal justice system.

4 It is important to note that while both political philosophy and neuroethics use ‘ideal’ and ‘non-ideal’ theories, their focus differs. Political philosophy views these theories through the lens of justice and societal structure. Neuroethics, however, shifts between theoretical ideals and the real ethical issues presented by neuroscience. This dissertation embraces these theories from a neuroethical standpoint, acknowledging their specific interpretations. Essentially, ‘ideal’ theory explores theoretical scenarios, while ‘non-ideal’ theory deals with the practical challenges we actually face. So, while both political philosophy and neuroethics employ the terms ‘ideal’ and ‘non-ideal,’ their

interpretations diverge based on context. Political philosophy anchors them around justice and societal constructs, while in neuroethics, they pivot between hypothetical musings and the palpable ethical challenges of neuroscience. Throughout this dissertation, I adopt these concepts as they resonate within neuroethics, recognizing their nuanced variations from traditional perspectives. In essence, while 'ideal' theory tilts towards hypotheticals, 'non-ideal' theory navigates the tangible complexities.

5 While the distinction between moral and normative judgments is a matter of ongoing philosophical debate, with various definitions offered, I choose to employ them in a specific manner within this context.

6 Therefore, I accept PEAs are justified in proceeding on the basis of the Punishment Claim. There is a rich body of debate in this field at the boundaries of neuroethics and penal theoretical analysis. It is acknowledged, but it falls beyond the scope of this dissertation.

7 As I will explain in later chapters, theorists such as Harris focus on neurointerventions in the context of 'neuroenhancement,' being the use of neurointerventions to enhance human capacities.

8 The United Nations had declared continuous solitary confinement beyond 14 days to represent cruel and unusual punishment, amounting to torture (UN General Assembly, *United Nations Standard Minimum Rules for the Treatment of Prisoners*: resolution / adopted by the General Assembly, 8 January 2016, A/RES/70/175).

9 *R v Capay*, 2019 ONSC 535.

10 In Canada, see decisions in *British Columbia Civil Liberties Association v Canada (Attorney General)*, 2019 BCCA 228; *Canadian Civil Liberties Association v Canada*, 2019 ONCA 243).

Although there have been measured shifts toward reform in the past three years in countries such as Canada (Mangat, 2018), it remains on the rise in the United States.

11 This issue has been discussed by various scholars (Boonin, 2008; Scheid, 1980; Zimmerman, 2011).

12 It does bear note that depending on the definition one adopts, it is open to question whether, sometimes, a neurointervention could be classified as punishment *at all*. For example, offering a neurointervention primarily aimed at rehabilitation might fail to represent punishment, given the absence of a connection between the offence and the punishment meted out (Bublitz, 2018, p. 311; Ryberg, 2020 Chapter 4; Sifferd, 2020).

13 I thank Jennifer Chandler for raising this point.

14 I am grateful to Ted McCoy for identifying this issue for consideration.

15 It is one of the more moderate incarceration rates in Western democracies but higher (or much higher) than in Western Europe.

16 This is beyond the costs of policing, investigation, prosecution, and ancillary court costs, which are, in themselves, exorbitant.

17 It bears note that of these two-thirds, half were reconvicted.

18 SUD is defined in terms of eleven criteria, including physiological, behavioural and cognitive elements, as well as consequences of criteria, any two of which qualify for a diagnosis. There is significant debate about the classification and definition of the disorder, which we will set aside for our purposes (Rehm et al., 2013).

19 It is worth noting that the specific emphasis on these aims can differ among various liberal theories or theorists, and there is presently no comprehensive liberal theory of punishment (Bedau & Kelly, 2019; Moore, 1997).

20 In line with other scholars, I consider this a valuable approach to addressing equivalency arguments, and I will delve into it further in subsequent chapters (Douglas, 2014c; Matravers, 2018, p. 88; Vallentyne, 2018b).

21 Rehabilitation is not solely a standalone theory of punishment but rather a feature encompassed by various comprehensive theories of punishment. While it aligns with consequentialist punishment by aiming for positive outcomes, it also contributes to limited retributivism by ensuring proportionality. Additionally, rehabilitation is central to communicative theories of punishment and plays a role in many liberal theories that seek to prevent future crimes by structuring social relations. However, for the present discussion, we can set this aside.

22 Thank you to Walter Glannon for identifying the importance of clarifying this point at this juncture.

23 I pause to note Douglas' vision of incarceration is one to which some believe we should aspire—and aligns with the underlying philosophy inspiring penal systems in certain Nordic countries, which I will briefly consider later (Davidson et al., 2000; Davidson & McEwen, 2012; Maier, 2020, p. 382; Nelken, 2009; Pereboom, 2018, p. 95; Pratt, 2007a, 2007b)

24 While my focus has predominantly been on compulsory neurointerventions, the significance of free and valid consent by a mentally competent offender cannot be understated. This factor plays a critical role when determining equivalence and permissibility. I extend my gratitude to Walter Glannon for highlighting the importance of clarification at this junture.

25 I am grateful to Jennifer Chandler for raising this important issue.

26 According to Persson and Savulescu, our primitive neurobiology hinders our ability to effectively address critical global challenges like climate change and the risks posed by weapons of mass destruction, thus jeopardizing our existence. They suggest that investigating innovative neurotechnologies, such as neurointerventions, offers potential as a solution to transcend these limitations (Persson & Savulescu, 2008, 2011a, 2011b, 2012, 2013; Savulescu & Persson, 2012)

27 It is important to note debates about our fitness for the future, particularly in the context of moral enhancement, raise pertinent questions when applied to criminal justice systems. These discussions, which often focus on our inability to overcome individual and local self-interests to avoid broader societal issues like the tragedy of the commons, bear similarities to the challenges faced in criminal justice. However, they also highlight distinct problems. In criminal justice, the issue is not just about overcoming self-interest but also about managing punitive responses that may no longer be suitable for the scale and complexity of modern societies. Addressing these debates within the criminal justice context underscores our struggle to adapt instinctive retributive responses to the complexities of modern legal systems—a transformation critical for ethical advancement. This transition is vital as we consider whether traditional punitive measures, often rooted in immediate emotional responses, can effectively address crime in diverse and extensive social settings where indirect consequences of punishment might be felt differently across the community. I thank Jennifer Chandler for suggesting this clarification.

28 Again, I thank Jennifer Chandler for her insightful feedback on this issue.

29 I thank Walter Glannon for recommending this clarification to ensure a nuanced view of the issue.

30 Moreover, poor nutrition, both prenatally and during childhood, is associated with an increased risk for risk factors for criminal offending and antisocial behaviours that manifest in childhood, teenage years, and later in life (Chew et al., 2018, pp. 25-26).

31 It is important to recognize that the distinctions in prevailing social and political realities among countries such as the US and other Western democracies are intricate and multifaceted (Barker, 2017; Shammas, 2014). While the Nordic models are often idealized, it is crucial to acknowledge that their realities are equally complex. However, in light of research in social and affective neuroscience that highlights the potential to shape and reshape socio-emotional skills associated with social and moral behaviour even in later stages of life, the Nordic models have proven to be a paradigm worthy of investigation in discussions surrounding penal reform (Davidson et al., 2000; Davidson & McEwen, 2012).

32 Or perhaps a supplement to.

33 For example, one of the leading objections to neurointerventions is their ability to circumvent rational capacities. In considering whether this is an 'in-principle difference,' Levy specifically dismisses this objection on the basis that "indirect subversion is an all too pervasive and all too powerful feature of the world we live in (Levy, 2020, p. 46). He argues that "far more injustice stems from indirect interventions than from direct, and this is not a situation that is likely to alter in the foreseeable future" (Levy, 2020, p. 33). We will build on this in what follows when we consider the concept of indirect subversion in exploring the parity principle.

34 As Levy states:

> It might be objected that in setting these matters aside, we set aside the ethics of neuroethics: the very heart and soul of the questions. There is some justice in this accusation: of course, it

> will be necessary to factor these concerns back into the equation in coming to an all-things-considered judgment of the advisability of using or promoting these drugs in actual circumstances. But clarity demands that we treat the issues raised by direct manipulation one by one, and that requires isolating them from one another, not conflating them. Moreover, in a book defending a conception of the mind as extended and knowledge acquisition as distributed, it is not special pleading to note that others – policy specialists, lawyers, sociologists and many kinds of medical professionals – are better placed than philosophers to analyze the issues set aside here. By focusing on the questions where I can best contribute, I hope thereby to advance the entire neuroethical agenda all the more effectively (Levy, 2007, p. 72).

35 Technically speaking, neuroscience is a field dedicated to the comprehensive study of the structure and function of the *nervous system*, with a primary emphasis on the intricate workings of the brain. There are various branches of neuroscience, including neuroanatomy –the study of the anatomy and function of the nervous system – and neurophysiology – the function of the nervous system. There are also dozens of subsets that include the fields of developmental, behavioural, cognitive, affective, systems, molecular, and computational neuroscience. In the clinical setting, advancements continue to provide a greater understanding of the healthy and pathological operation of the brain – how the brain operates in sickness and health. This can bear upon diagnoses and treatment for diseases and, in some instances, may offer predictive information before the onset of symptoms (see C. M. Altimus et al., 2020; Hall et al., 1993; Stender et al., 2014)In recent years, neuroscience has become heavily interdisciplinary, drawing from many disciplines, including psychology, medicine, genetics, mathematics, linguistics, engineering, and chemistry (C. M. Altimus et al., 2020; Bear et al., 2020;

Lipina, 2014). This has led to many specialized fields, such as cognitive, affective, behavioural, developmental, evolutionary, computational, and, more recently, theoretical neuroscience (Coward, 2013).

36 This dissertation falls within this field. It focuses on the intersection between neuroscience and ethics in the emerging field of 'neuroethics' (Roskies, 2002; Roskies, 2020a) and the intersecting field of 'neurolaw' (Chandler, 2018; Nicole A. Vincent et al., 2020).

37 This traces a trend toward research in the field of 'neuromarketing,' which uses neuroscience and brain imaging technologies to understand consumer preferences and purchasing patterns. By 2010, it was estimated that over 300 commercial organizations were already involved in the study in this area (Ulman et al., 2015).

38 A great deal of evidence has demonstrated that people are very poor at detecting falsehood, with accuracy rates little better than chance (Delgado-Herrera et al., 2021; Seaman, 2008). As such, the potential utilization of fMRI lie detection in the legal context, particularly the criminal context, has been the subject of great scholarly debate in the literature. However, based on shortcomings in these technologies, there are many who believe that "[a]t present, there is no question but that neuroimaging lie detection is nowhere near ready for courtroom (or any other forensic) use" (Delgado-Herrera et al., 2021; Farah et al., 2014; Hsu et al., 2019; Seaman, 2018, p. 210)

39 It is important to note, however, that some studies challenge this view by showing that neuroimages might not have as much effect as once thought (see, for example, Bennett & McLaughlin, 2023).

40 A term borrowed from (Ryle, 1949).

41 I acknowledge the quote “Truth is stranger than fiction” is often attributed to the American writer and humorist Mark Twain (also known as Samuel Clemens)—although its exact origin is unknown.

42 Which support neurons but do not produce electrical impulses. The primary purpose of neurons and glia is to facilitate the flow of energy and information within the brain.

43 Neurons transmit electrical impulses, known as ‘action potential,’ down the cell axons. This, in turn, releases a neurotransmitter at the end of the neuron, known as the ‘synapse.’ The synapse forms the connection that links neurons together.

44 Assuming the current format of this dissertation.

45 Moreover, because each cell has a constellation of interconnections, the activation of one neuron can influence the activation of an average of ten thousand other neurons. This comprises a ‘neural net profile,’ which creates a pattern of neural activity clustered into a functional whole.

46 ‘Energy’ denotes the metabolic processes essential for neuronal activity, while ‘information’ refers to the encoding and transmission of signals—processes enabled by these metabolic activities.

47 The brain receives environmental inputs via distinct sensory systems. These inputs are then ‘routed’ to specific brain networks for processing and consolidation. For instance, visual inputs are processed in the visual cortex, auditory inputs in the auditory cortex, somatosensory inputs in the somatosensory cortex, and vestibular inputs in the vestibular nuclei and related areas.

48 This builds on a conceptual framework often used in cognitive science and related fields to describe different perspectives on the relationship between cognition and the environment.

49 Integration allows for bidirectional influence between the brain and bodily systems, as well as with genes influenced by epigenetic and environmental factors (Gintis, 2000; Menzel Jr, 1974; Siegel, 2012, p. 27; Walker, 2009, p. 31).

50 Memory, a key function of the human brain, is typically divided into several types. The two main categories are declarative (explicit) memory and non-declarative (implicit) memory. *Declarative Memory*: This type of memory involves information that can be consciously recalled and verbally described. It is further divided into two subcategories.

a. Episodic Memory: These are memories of specific personal experiences or events and their context (time, place).

b. Semantic Memory: This comprises factual information, general knowledge about the world, and language.

Non-Declarative Memory: These are memories that influence behaviour without conscious awareness. They cannot be consciously accessed or verbally described. Non-declarative memory is further divided into several types (Squire, 2009; Zawadzki & Adamczyk, 2021)

51 A more centrally located 'limbic system' integrates and distributes information between these two brain regions—lower and higher cortical structures (Urban & Rosenkranz, 2020). This might be better understood as 'limbic *systems*' (Rolls, 2015).

52 An appreciation of which is necessary to avoid the 'fallacy of localization.'

53 It is important to note that based on advancements in neuroimaging, particularly resting-state fMRI (rs-fMRI), there are advancements toward understanding the functional organization of consciousness across multiple scales that seem necessary for consciousness (Huang et al., 2018). But the identification of robust biomarkers for consciousness is still lacking (Campbell et al., 2020), as is a consensus on a comprehensive *theory* of consciousness (Kent & Wittmann, 2021; Northoff & Lamme, 2020).

54 The trolley problem is a series of thought experiments in ethics and psychology involving stylized ethical dilemmas of whether to sacrifice one person to save a more significant number – reflecting the classic distinction between utilitarian or 'consequentialist' moral judgements (sacrifice one to save others) and deontological judgement (certain acts, such as killing, are always morally impermissible).

55 Discussion of these experiments (as all those in this Chapter) are all subject to the qualifications explored in this chapter, including the limits inherent in present technologies, and inherent empirical and conceptual issues.

56 In the criminal justice context, theorists sometimes adopt or use interchangeable or reference similar terms such as direct brain interventions (Craig, 2016; Greely, 2012; Shaw, 2012; Vincent, 2014), biological interventions (Olivia Choy et al., 2018), neurobiological interventions (Specker et al., 2017), neuromodulation (Glannon, 2015; Schermer, 2015), and more recently, 'crime preventing neurointerventions (David Birks & Thomas Douglas, 2018). The latter, specific to discussions in the criminal justice system, has been defined as "interventions that exert a physical, chemical, or biological effect on the brain in order to *diminish the likelihood of some forms of criminal offending*" (D. Birks & Thomas Douglas, 2018, p. 2 emphasis added).

The term 'neurointervention' is reframed within the 'neuroenhancement debate,' which primarily addresses ethical issues concerning the use of novel neuro-technologies for 'enhancement' rather than 'treatment' (Buchanan, 2011; Focquaert & Schermer, 2015, p. 139). It involves the application of such interventions outside the criminal justice system, extending to academics, the workforce, and sports. (Holm & McNamee, 2011; Tamburrini & Ta¨nnsjo€, 2011) (Bruhl et al., 2019). It often raises contentious debates about the distinction between 'treatment' and

'enhancement' and includes discussion around 'moral enhancement', which pertains to biomedical interventions aimed at improving moral reasoning for the benefit of individuals, and humanity generally (Carman, 2021; Earp et al., 2018; Harris, 2011, 2014a; Persson & Savulescu, 2008, 2011a, 2011b, 2012, 2013, 2015; Raus et al., 2014; Savulescu & Persson, 2012; Sparrow, 2013; R. Sparrow, 2014; R. J. Sparrow, 2014; Wiseman, 2016).

However, 'enhancing' moral capacities has sometimes been discussed in the criminal justice system context (Earp et al., 2018, p. 167; Persson & Savulescu, 2012, p. 402). Technologies aimed at 'enhancing the capacities of criminal offenders for moral reasoning. Technologies aimed at 'enhancing the capacities of criminal offenders for moral reasoning. Leaving this aside, I will now adopt the concept of 'neuroenhancement' to discuss the class of interventions considered in this debate. I will adopt the concepts of 'cognitive neuroenhancement' and 'moral neuroenhancement' as analogues for these specific subclasses. It is sufficient to say that many technologies we consider below have been considered outside of the criminal justice debate and for ethical issues about improving humans generally. While we will discuss neurointerventions primarily in the concept of the criminal justice system, it is essential to remember that these technologies need not be limited in their use to criminal offenders or the criminal justice system.

57 For example, for hundreds of years, occupants of the Amazon basin used ayahuasca, a brew derived from the Banisteriopsis caapi and Psychotria Viridis, which contains MAO inhibitors and N-dimethyltryptamine or DMT.

58 Although we will discuss procedures involving entering the skull, we will not extensively explore psychosurgeries, which involve surgical intervention or removal of brain portions like tumors. It is

improbable that mandatory techniques aimed at treating criminal offenders or moral defects would be considered, at least in the foreseeable future.

59 The increasing off-label use of these cognitive enhancers particularly in academic and workplace settings is a matter that has been widely discussed in the context of the neuroenhancement debate (Bruhl et al., 2019; Caviola & Faber, 2015; Franke et al., 2014; Lucke & Partridge, 2012; Maturo, 2012; Sahakian & LaBuzetta, 2013; Sebastian & Sahakian, 2018; Turner & Sahakian, 2006).

60 While antiandrogens do not strictly fall under the category of neurointerventions since they target the endocrine system, their indirect impact on the brain through reducing sexual desires warrants consideration.

61 tDCS has also attracted attention in the neuroenhancement literature, as it has been suggested to enhance performance in several cognitive domains. This has led to the emergence of consumer tDCS devices, raising ethical concerns and prompting discussions within the scientific and medical communities (Luber & Lisanby, 2014; Wexler & Reiner, 2018, p. 272).

62 TMS involves electromagnetic induction to induce electrical currents and has shown efficacy in treating various neurological and psychiatric disorders. Repetitive TMS (rTMS) is a recent variation that produces lasting effects beyond the stimulation period. On the other hand, tDCS utilizes a low electrical current generated by a battery-powered apparatus to target specific brain areas. It has been explored for enhancing cognitive functions and treating conditions such as Parkinson's disease, Alzheimer's disease, depression, and schizophrenia

63 Ethical concerns regarding DBS vary depending on whether it utilizes open-loop or closed-loop stimulation. Open-loop DBS delivers fixed electrical impulses without monitoring brain activity,

whereas closed-loop DBS involves real-time monitoring and adjusts stimulation accordingly (Goering et al., 2017; Parastarfeizabadi & Kouzani, 2017)

64 Classifying them as a 'subtype of neuroprosthesis.'

65 As discussed in various studies (Chew et al., 2018, p. 17 see also contra; Gagne, 1981; Giltay & Gooren, 2009; Grasswick & Bradford, 2003; Keating et al., 2006; Laschet & Laschet, 1975; see generally Ryan, 2020, pp. 285-285).

66 Special thanks to Professor Jennifer Chandler, co-author of the referenced study, for her invaluable insights into this area of research.

67 Again, thanks to Jennifer Chandler for bringing these studies to my attention.

68 William Bülow discusses this in the context of communicative theories of punishment, arguing that restoring cognitive capacities would permit offenders to achieve penological aims through a form of 'secular penance' (Bülow, 2020).

69 Thanks to Jennifer Chandler for this insightful example.

70 Classifying them as a 'subtype of neuroprosthesis.'

71 with high accuracy using a BTBI connection involving EEG and focused ultrasound

72 By implanting a microstimulator and utilizing EEG and a BCI to transmit Bluetooth-emitted signals.

73 Human genetic engineering (HGE), involving the alteration of human genes, has been a topic of considerable debate, with its historical roots dating back to early ideas of selective breeding and eugenics. Recently, genome-editing technologies, most notably CRISPR-Cas9, have developed, enabling site-specific manipulation of the genome in somatic (non-reproductive) cells (Canli, 2015; Jinek et al., 2012). For example, it has recently been suggested that, in theory, HGE and gene editing

could target specific domains, enhance traits such as empathy, reduce disposition for moral aggression, and bestow capacities for complex moral reflection (Rakic, 2019). However, at least for the foreseeable future, such applications are speculative, given our current technology and understanding of the complex human genetic structure. Further, some have suggested that trying to enhance one specific trait might result in "unforeseeable, unintended harms because of the complexity of human genetic structure" (Fukuyama, 2003, p. 74).

74 Assuming each multiplication operation takes a nanosecond (one billionth of a second), even performing these calculations would require an exorbitant amount of time, several orders of magnitude longer than the age of the universe itself.

75 These figures assume the average connectivity and features of the brain in their entirety while, admittedly, there would be variation across various brain regions scanned or mediated.

76 Indeed, it is likely neuroimaging techniques will continue to improve, and forms of neuromodulation continue to increase in precisions—as illustrated in optogenetics. But at present, these limitations cannot be overlooked.

77 A saying often attributed to Albert Einstein.

78 Subsequently, it appears, alongside others, that they provide at least some reasons to be suspicious of our deontological judgments and interpret his work as lending credence to utilitarian theories (Greene, 2003; Greene, 2008, 2013; Singer, 2005).

79 While the exact role of this treatment in his subsequent death remains a topic of debate, it is widely acknowledged that it had significant detrimental effects on his physical and mental well-being. In contemporary times, the criminalization of personal drug possession in certain Western countries has sparked debate. Some argue that by adopting a 'moral' rather than a 'medical' model of addiction,

these policies have faced criticism for exacerbating the public health crisis of drug addiction and contributing to a rise in tragic deaths caused by toxic drug overdoses. This perspective involves normative claims and raises complex issues open to discussion, although there are indications of a shift towards decriminalization (Russoniello et al., 2023).

80 I am grateful to Jennifer Chandler for identifying this important issue.

81 Levy calls them 'Internal interventions,' while others adopt different definitions. I use 'direct interventions' for the purpose of discussing the parity principle.

82 Levy uses the term 'external interventions,' while others reference 'environmental interventions' (Douglas, 2014d, p. 216; Pugh, 2019, p. 84). I will use the term 'indirect interventions' here, as it is most commonly used in the literature (Bublitz & Merkel, 2014; Craig, 2016; DeGrazia, 2014; Focquaert & Schermer, 2015; Greely, 2012; Harris, 2014b; Raus et al., 2014; Shaw, 2012).

83 The question of what is the 'mind' in an ontological sense is an issue we cannot discuss in depth. It is sufficient to identify there is some question about what should be conceived as the 'mind'—what sorts of carriers or vehicles it could comprise. Theorists have proposed constraints to ensure the EMT does not become "like functionalism, too liberal in its ontology of the mind" (Heersmink, 2016; Heinrichs, 2018). We can set this aside for the time being.

84 Levy adopts two forms of the parity principle. This is the 'weak version'. He also proposes a stronger version that states as follows:

> EPP (strong): Since the mind extends into the external environment, alterations of external props used for thinking are (ceteris paribus) ethically on par with alterations of the brain (Heinrichs, 2018, p. 63; Levy, 2007).

85 A this point, I reiterate the parity principle, building on the ‘in-principle’ constraint discussed earlier, which assumes the safety and effectiveness of neurointerventions (Levy, 2007, p. 73). Levy’s precise terminology is to say ethically relevant differences are those that remain sound regardless of “how much the technologies improve” or the “political and social context [in which] they are developed” (Levy, 2007, p. 73)So we again temporarily disregard the multitude of practical issues previously outlined, issues that pose clear and present risks for any implementation of neurointerventions in the real world and our current punishment practices.

86 This keen insight regarding the distinction between conscious awareness and control, provided by Jennifer Chandler, PhD Examiner, challenges the often implicit assumption that these two aspects necessarily coincide. Chandler highlights the potential for individuals to be aware of an intervention’s effects without possessing the capability to resist or alter those effects. This observation suggests that the ethical landscape of neurointerventions is more complex than typically acknowledged, where conscious awareness does not always grant control.

To explain further, Chandler’s perspective introduces a valuable dimension to the debate on neurointerventions, particularly in how we conceive of mental freedom and autonomy. If individuals can be aware of but powerless over the effects of an intervention, this might constitute a unique kind of infringement on autonomy that differs from situations where individuals are either unaware or fully in control. This insight necessitates a more nuanced discussion within neuroethics, one that considers the permutations of awareness and control as distinct yet intersecting factors influencing the ethical evaluation of neurointerventions. Further attention to this nuanced aspect exceeds the scope of this dissertation but represents a critical area for future research.

87 This is not necessarily the case, for example, to the extent that neural prosthetics, such as DBS, allow input from the user, while other interventions, as I will explain below, also have 'composite' elements.

88 Levy discusses his parity principle in the context of cognitive enhancement as part of the 'bio enhancement debate.' However, I will argue these considerations apply with equal, and perhaps even greater force to issues about criminal justice practices.

89 Referencing the title of Timothy Wilson's book "Strangers to Ourselves: Discovering the Adaptive Unconscious (Wilson, 2002, p. 98).

90 Joseph Campbell describes this in the context of archetypical comparative mythology.

91 Panksepp proposes six systems functionally dedicated to social interactions, including systems related to fear, rage, lust, care, grief, play, and seeking (Davies, 2020, p. 326; Panksepp, 2004, 2012).

92 However, while the autonomic and emotional processes are generally considered 'lower-level' and somewhat autonomous, they can be influenced to varying extents by higher cognitive functions, reflecting a degree of 'top-down' control. However, the capacity for such control is typically limited and varies among individuals and contexts. Understanding these processes and how they can be modulated is a complex task that remains an active area of neuroscience research.

93 For example, in one study, subjects evaluated two police chief candidates—one male and one female, with one being 'streetwise' and the other well-educated. Regardless of the presented characteristics, the male candidate was preferred, with justification conveniently fitting his profile in each case. This preference was not consciously based on gender, but instead, their subconscious bias tailored the job requirements to favour the male candidate. This reveals a confabulated criterion for qualification, where the underlying bias was not consciously recognized (Uhlmann & Cohen, 2005).

94 While acknowledging the reciprocal influence between the mind and the environment, for the sake of simplicity, I will set aside this complexity and separate the two.

95 Identification and discussion of certain of these studies can be attributed various other theorists (Davies, 2020; Kahneman, 2011; Levy, 2020; Ryberg, 2020)

96 Although the extent of conscious control over these autonomic functions can vary.

97 Like activating brown fat tissue. Brown fat tissue, unlike the regular white fat, is packed with mitochondria, the cell's powerhouses, which burn calories to produce heat. Instead, it was was primarily driven by controlled, forceful breathing producing heat.

98 Its single-subject design hampers generalizability. The observed brain activation may not infer causation, but merely correlation, possibly falling prey to the localization fallacy. The used methods (PET/CT, fMRI) are powerful but have inherent spatial and temporal resolution limits, and rely on blood flow proxies instead of direct neural activity. Lastly, Hof's forceful breathing could confound the effect of top-down control, adding complexity to disentangling these influences. These aspects urge careful interpretation of Hof's case and its implications for top-down control.

99 Meditation and mindfulness have rich cultural traditions and come in many forms. For example, the Buddhist tradition has refined methods for stabilizing attention, while Hindu and Indian practitioners focus on cultivating sustained and steady attention. Mindfulness and meditation have gained popularity in popular science and self-help books, which can lead to sensationalization. It is important to approach this topic with caution due to the novel nature of the subject, the influence of popular culture, and the potential for exaggerated claims. In the following discussion, I will strive to avoid sensationalization and present a balanced view.

100 Subject to the issues we identified in the last Chapter, such as localization and causation/correlation.

101 While external stimuli can shape self-organization to some extent, the brain's inherent spontaneity and intrinsic dynamics are the primary drivers of this phenomenon.

102 This builds on a conceptual framework often used in cognitive science and related fields to describe different perspectives on the relationship between cognition and the environment.

103 For discussions about the concept of 'community of minds' and 'cultural cognition' see: (Hutchins, 1995; Surowiecki, 2005).

104 The concern is that the EMT's boundary determination, when pushed to an extreme, resembles functionalism, which identifies mental states by their functional role, not by physical or biological aspects. It thus opens the possibility of various entities, including brains, computers, or a system of cultural artifacts, possessing a "mind." Critics, however, perceive functionalism as excessively liberal in its understanding of the mind's ontology, which concerns the nature of existence and reality. Despite the challenges, the criteria for defining the mind's boundaries remain a contentious issue in ongoing theoretical and empirical debates (Adams & Aizawa, 2001; Heersmink, 2016; Heinrichs, 2018)

105 Again, I would note that while this thesis primarily focuses on the examination of neurointerventions in the context of criminal justice, it is important to acknowledge that the topics and issues discussed intersect with the broader debate in contemporary neuroethics of moral neuroenhancement. These discussions have been extensively explored by scholars, and I will draw on similar themes in what follows (Bruhl et al., 2019; Buchanan, 2011; Carman, 2021; Earp et al., 2018; Focquaert & Schermer, 2015, p. 139; Harris, 2011, 2014a; Holm & McNamee, 2011; Persson

& Savulescu, 2008, 2011a, 2011b, 2012, 2013, 2015; Sahakian & LaBuzetta, 2013; Savulescu & Persson, 2012; Sebastian & Sahakian, 2018; Sparrow, 2013; R. Sparrow, 2014; R. J. Sparrow, 2014; Tamburrini & Ta¨nnsjo€, 2011; Wiseman, 2016).

106 Known as the 'principle of alternative possibilities' *PAP* (Frankfurt, 1969, 1988b; Haji, 2009).

107 In alignment with this species of thought, Harry Frankfurt presented a unique thought experiment. Imagine a person with a computer chip implanted in their brain without their knowledge. This chip could compel them to act in a specific way. However, suppose the individual independently decides to perform the same action that the chip was designed to enforce. In that case, Frankfurt argues that they should still be considered morally responsible for their choice, even though they technically could not have *acted* differently. This provocative scenario calls into question our notions about the prerequisites for moral accountability. It suggests that making an independent choice, even in a constrained context, is sufficient for moral responsibility. This distinction and these thought experiments have been widely debated in many cases through a series of reformulated thought experiments (Fischer, 1998; Widerker & McKenna, 2003). I will not pursue these lines of inquiry here.

108 It is worth mentioning again that the boundaries between these domains are not always clear, and the categories we propose should be viewed as guiding tools rather than rigid classifications.

109 Earlier, we even noted evidence that the use of psychotropic agents to alter cognition may in fact date back to prehistoric times and may have even played a role in our evolutionary history (Earp et al., 2018, p. 175; Homan, 2011; McKenna et al., 1984; Merkel et al., 2007, p. 11; Merlin, 2003; Rodriguez Arce & Winkelman, 2021; Wolpe, 2018, p. 220).

[110] I thank Jennifer Chandler for this insight. As she also correctly observes, such interventions may also precipitate a reconsideration of whether the justice system's primary aim is indeed moral blame or if it should focus more on public safety and incapacitation. This signals a question of whether a shift could allow reason-bypassing interventions, accepted as self-control mechanisms by the individual, to remain within the ethical domain of personal responsibility and moral evaluation.

111 Because the mind supervenes on the brain, there is "a continuous transaction between current states of the brain, body, and the environment" (Clark, 2008a; Pouw et al., 2014, p. 53).

112 One of the more troubling issues is the requirement of *valid* consent for biomedical procedures, precipitating a 'deeply polarized' debate in discussions of neurointerventions as criminal punishment (Beauchamp, 2013, p. 59; Bomann-Larsen, 2013; Caplan, 2008, p. 103; Douglas et al., 2013; W. Green, 1986, pp. 16-17; Kutcher, 2010a; Manson & O'Neill, 2012; McMillan, 2014; Pugh, 2018; Ryberg, 2012; 2020, p. 26; Ryberg & Petersen, 2011; Shaw, 2012; Sifferd, 2020; Stefano, 2021, p. 214; Vanderzyl, 1994). I set this aside here as the focus is on *compulsory* interventions, maintaining a skeptical stance on the feasibility of achieving valid consent in most cases under current penal practices..

[113] *Canadian Charter of Rights and Freedoms*, s 7, Part 1 of the Constitution Act, 1982, being Schedule B to the Canada Act 1982 (UK), 1982, c 11 a

[114] Universal Declaration of Human Rights, GA Res 217A (111), UNGAOR, 3d Sess, Supp No 13, UN Doc A/810 (1948) 71 [UDHR].

[115] International Covenant on Civil and Political Rights, 16 December 1966, 999 UNTS 171 (entered into force 23 March 1976) [ICCPR],

[116] This discussion was informed by contributions from Professor Greg Hagen of the examination committee, who emphasized the importance of exploring these issues—a sentiment I share in respect to an area I hope to explore in great depth in future research. Further legal aspects identified, but not fully explored here but acknowledged include not only freedom of conscience, life, liberty, but security of the person, and protection against cruel and unusual punishment as outlined in sections 7, and 12 of the Charter, respectively.

[117] I thank Jennifer Chandler for this insightful observation.

[118] Jennifer Chandler raises an important concern regarding the practical effectiveness of rights to mental integrity within a system potentially coercive enough to induce individuals to request neurointerventions. This concern highlights the potential for a system where individuals may seek neurointerventions not entirely voluntarily but as a less detrimental alternative within the constraints of the criminal justice system. While this dissertation suggests recognition of a right to mental integrity as a potential safeguard, a full articulation or defence of such a right is far beyond the scope of discussion here. Chandler's insights suggest a need for further discourse on strengthening these rights to adequately address such complex scenarios of consent.

119 This is a reference to description of such a right by Jans Bubliz (Bublitz, 2020a).