Free Will, Causality, and Neuroscience

Value Inquiry Book Series

Founding Editor

Robert Ginsberg

Executive Editor

Leonidas Donskis†

Managing Editor

J.D. Mininger

VOLUME 338

Cognitive Science

Edited by

Francesc Forn i Argimon

The titles published in this series are listed at *brill.com/vibs* and *brill.com/cosc*

Free Will, Causality, and Neuroscience

Edited by

Bernard Feltz
Marcus Missal
Andrew Sims

BRILL
RODOPI

LEIDEN | BOSTON

Cover illustration: image by Gerd Altmann from Pixabay.

The Library of Congress Cataloging-in-Publication Data is available online at http://catalog.loc.gov

Typeface for the Latin, Greek, and Cyrillic scripts: "Brill". See and download: brill.com/brill-typeface.

ISSN 0929-8436
ISBN 978-90-04-37291-7 (hardback)
ISBN 978-90-04-40996-5 (e-book)

This book is printed on acid-free paper and produced in a sustainable manner.

Contents

PART 3
Causality and Free Will

Acknowledgements

The Authors thank the *Actions de Recherches Concertées* (ARC, UCLouvain), and the *Fonds pour la Recherche Scientifique* (FRS/F.N.R.S, Belgium) for supporting the research project *Free Will and Causality* which has led to this book.

The Authors

Bernáth, László, Bernáth, László, Hungarian Academy of Sciences and Eötvös Loránd University, Institution of Philosophy, Budapest
Bonicalzi, Sofia, Ludwig-Maximilians-Universität München, Fakultät für Philosophie, Wissenschaftstheorie und Religionswissenschaft
Corey, Hill Allen, Georgia State University
Drozdzewska, Anna, Université catholique de Louvain, Institut supérieur de philosophie
Feltz, Bernard, Université catholique de Louvain, Institut supérieur de Philosophie
Loveall, Bradley, Georgia State University
Mele, Alfred, Florida State University
Missal, Marcus, Université catholique de Louvain, Institute of Neurosciences
Nahmias, Eddy, Georgia State University
Schlosser, Markus, University College Dublin
Sartenaer, Olivier, Universität zu Köln
Sims, Andrew, Université catholique de Louvain, Institut supérieur de Philosophie

Introduction

Bernard Feltz, Marcus Missal and Andrew Sims

What is the relationship between mind and body? Even in antiquity, the contrast between Platonic and Aristotelian conceptions of the relationship between soul and body anticipate this anthropological debate, which has recently been taken up under the banner of a "naturalisation of consciousness." In the modern period, the dialogue between Descarte's neo-Platonism and Spinoza's monism—and later, the Kantian critique—can be interpreted as a continuation of these themes. The passionate discussions in the 20th century between Sartre and Merleau-Ponty about phenomenological "Being-in-the-World" offer a new approach to that which is specifically human. At the heart of all these traditions, language and the capacity for meaning-making play a decisive role. Paradoxically, it is at the very moment when philosophy discovers the corporeal dimension of humanity—in dialogue with the natural sciences, furthermore—that the ubiquity of language in all behaviour is freshly brought into view. Paul Ricoeur and Jürgen Habermas are exemplary of such a position. A human being is a corporeal entity but human behaviour is unintelligible in abstraction from the role of language.

Recent developments in neuroscience have led to a fresh perspective on this set of issues. The relationship between mind and body is not only a philosophical matter. A close dialogue with the experimental sciences is not only possible but indeed also necessary. In recent decades, new methods for studying the brain have produced important theoretical advances and led to the rapid development of various disciplines, among which are neuroscience and the cognitive sciences. The philosophy of neuroscience has itself grown considerably. In this context, one influential line of thinking tends toward the thesis that free will is pure illusion and that the principle of causation in all its rigor leads inexorably to the rejection of a concept of free will likely to contribute to an understanding of human behaviour.

This conducts us to propose two introductory developments : free will and causality.

1 Free Will

There are several epistemological positions that yield different conceptions of free will. Beginning from his work in cognitive psychology, Daniel Wegner

- 978-90-04-40996-5

(2002) constructs what he calls a "theory of apparent mental causation" which explains the conscious experience of our own causal efficacy during action as a post-hoc reconstruction. On the basis of Libet's (1985) experiments, he argues that the conscious experience of acting is pure illusion. It is language—in an a posteriori reorganisation— that inscribes a particular action in the autobiographical memory. The subject then acquires an illusory sense of herself as the author of that action.

In the same line, in the course of a discussion of theories of self-organisation, decision theory, and cognitive models, Henri Atlan (2011) proposes a neo-Spinozist interpretation of human behaviour which is characterised by its total determinism. For him there is a kind of freedom that is linked to the impossibility of predicting behaviour, but this impossibility is purely epistemological. Human behaviour is deeply determined and freedom consists in making this determination our own. In this context, free will proper has no place and is qualified as a "necessary illusion" in that, while we must act as if our decisions were efficacious, this efficacy is fictive on account of the principle of total determinism. Both of these perspectives just discussed are characterised by the conception of a deterministic world and the conception of scientific explanation in terms of causation. Free will is understood as calling into question a principle of causal continuity that these conceptions imply.

This introduction is not the place for a detailed discussion of each of these sorts of models. Simply put, our working hypothesis is that the concept of free will remains pertinent without coming into conflict with any scientific practice that requires the concept of causation in its explanations. Language, as an emergent process of brain activity, contributes to the development of behaviour which count as unified actions taking place over long intervals of time and which makes it possible to conceive of an acting body which is both corporeal and yet aptly described as "free."

More precisely, intentional action in human beings is a function of both the initial conditions of the distributed neural networks that are involved (that is, the "brain state": emotions, physiological state, autobiographical and implicit memory) and the circumstances and events at some moment (rest, activity, social interactions, etc…). The expression "free will" is used in order to describe this situation of interaction between an agent whose nervous system reacts to a prior set of events according to both its state at that point and according to an intentional logic. Intentional action refers explicitly to an operation that involves the capacity to represent a future state of the world. The concept of free will refers both to behaviour that has this intentionality and for which it is not entirely determined. Our hypothesis is that the use of language is an activity that allows such an operation.

To defend this conception, we have to analyze more precisely the relation with causality.

2 Causality

From the philosophical point of view, the study of the concept of causality started with Aristotle (384 BC–322 BC) who proposed four different types of causes (material, formal, efficient, final; *Physics* II and *Metaphysic*s V 2). During the following centuries, the concept of causality has continued to be interpreted in Aristotelian terms. David Hume initiated the modern approach of causality. He recognized the importance of causal beliefs for human understanding. However, he convincingly demonstrated that causality itself is not observable. Describing colliding objects, David Hume wrote: "When we consider these objects with the utmost attention, we find only that the one body approaches the other; and the motion of it precedes that of the other without any sensible interval" (Hume 1739). The argument of David Hume seems logically impossible to contradict: a necessary connection between events cannot be observed or measured. Only contiguity and succession can be observed. Causality seems indispensable to human understanding but could not be founded rationally and causal inferences are made on the basis of non-causal co-variations. If we follow David Hume's philosophy, the mind is a white sheet of paper and only learned associations can form the base of human knowledge. Immanuel Kant considered Hume's conception of causality as deeply unsatisfactory. In Kant's approach, causality is an a priori category of understanding, a logical necessity for the possibility of experience. Categories of understanding are *a priori* features of the mind. Therefore, for Immanuel Kant, the mind is not a white sheet of paper. The British philosopher Bertrand Russell tried to clause the debate by declaring the concept of causality obsolete: "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm" (Russell 1912). However, we suggest that simply giving up the concept of causation at the macroscopic level is unsatisfactory. More specifically, the concept of causation is central to the notion of free will. Indeed, free decisions could cause behavior if humans enjoy free will and this question is central in modern philosophy.

The modern concept of causality has been deeply influenced by physics and psychology during the XXth century and has a deep impact on causality in neurosciences.

According to the physicist Max Born (1949): "Causality postulates that there are *laws* by which the occurrence of an entity B of a certain class depends on

the occurrence of an entity A of another class, where the word 'entity' means any physical object, phenomenon, situation, or event. A is called the cause, B the effect." The concept of "lawlike" necessity is important in the contemporary approach to causation. Moreover, Max Born added that the cause should precede (or at least be simultaneous with) the effect and that there must be some sort of spatial contact between the cause and effect (even if it is by way of a chain of intermediaries). Albert Michotte (1949/1963), profoundly influenced by Immanuel Kant, considered the possibility that humans actually "perceive" causality directly through the activation of an encapsulated specific brain detector receiving a particular pattern of spatio-temporal inputs (Wagemans et al. 2006). Michotte used abstract visual stimuli, such as shapes that moved and collided in various ways, and made detailed manipulations of their spatial and temporal properties. His subjects responded with verbal descriptions of the resulting "scenes," and Michotte determined whether they thought there was a causal percept ("object A caused object B to move") or not. Michotte concluded that humans perceive causality as a Gestalt, similar to the way they perceive shape, motion, or other fundamental qualities in the world.

Michotte's results have been replicated in contemporary experiments (for review, see Scholl and Tremoulet 2000). Whatever their interpretation, Michotte's experiments and those of his followers clearly show the prevalence of causal judgment in psychology and behavior (Badler et al. 2010; Badler et al. 2012). Suggesting that causality is an illusion is epistemologically counterproductive. Similarly, idea that causal beliefs are elaborated on the basis of passive observations of covariations suffers from obvious limitations. Indeed, readings of a drop of atmospheric pressure on a barometer covaries with storms occurrence. However, nobody will claim that manually changing the reading of a barometer could cause a storm. Genuine causation must be distinguished from spurious. The modern approach to causality inference that is emerging can be thought of in terms of graphs and probabilities. The fundamental idea is that a cause raises the *probability* of occurrence of an effect. Making causal hypotheses is very similar to elaborating a scientific theory from experimental data (Glymour 2001).

More recently, any works are specifically oriented to causality in neuroscience. Craver distances himself from "law-like" necessity causality and defends a mechanistic conception of causality. To explain is to show multilevel mechanisms conducting the transition from state 1 to state 2. In the same line, Woodward proposes an interventionist concept of causality where the articulation between levels of organization in the brain is essential.

These diverse conceptions of causality are present in this book. Each author dialogues with one or other conception in order to think the possibility of free will.

3 Content

The relation between free will and causality is an important focus of this book. That is why we organize it into three parts. In the first part, "Intention and Consciousness,", the objective is to consider a priori theories of the meaning of intentional action in light of our increasing knowledge about the architecture of cognition, and to probe intuitive ideas about the relationship between control, intention, and consciousness. The compatibility of intention with efficient causality is also analysed. In the second part, "Libet-Style Experiments," while Libet's famous experiments are generally considered as defending a causality which reject free will, we would like to reconsider these experiments in light of the variety of ways in which they have been instantiated as well as the sorts of theories which they are intended to refute. The third part, "Causality and Free Will,", aims to clarify the ways in which language has an impact on human behaviour, and in the way that it allows a rich scope of flexibility and planning that would otherwise be out of reach. The relation with causality is the main topics, first in articulation with mental causation, finally in the context of emergence.

Specifically, in "Perceptual Decision-Making and Beyond," Andrew Sims and Marcus Missal extend models of perceptual decision-making in psychophysics in order to elaborate a theory of intentional action that does not rely on the propagation of content from abstract propositional attitudes to sensorimotor representations in the concrete moment of action (e.g. Pacherie 2008). Instead, this model conceptualises intentional action as a process in which quasi-perceptual representations bias the evolution of a "decision variable" into a state space which represents the sensorimotor consequences of a particular outcome. This is intended to be an alternative to the sort of causal theory of action that links action to causation by propositional attitudes.

Markus Schlosser is more optimistic that the traditional picture of action and control can be retained, and illustrates this by walking us through a challenge posed to that picture by dual-process theories of cognitive architecture. In his piece "Dual-System Theory and the Role of Consciousness in Intentional Action" he carefully distinguishes between various kinds of control and guidance, and concludes that the traditional picture can be preserved given qualifications about the role of consciousness.

Nahmias, Allen, and Loveall ask their participants the question: "When do Robots have Free Will?" They do so in order to further probe the importance that attributions of phenomenal consciousness have to ascriptions of free will. Their guiding hypothesis is that phenomenal consciousness matters because for an agent to be free and responsible requires that agent to care about one's choices and their consequences, and that care requires the capacity to feel emotion. Their results provide tentative support for this hypothesis.

The second part of the book contains two chapters that revisit themes in the empirical literature. In his "Free Will and Neuroscience," Alfred Mele considers Libet-style arguments against free will in light of recent updated instances of these studies. He considers two specific arguments for the nonexistence of free will that he takes to be refuted and concludes that recent studies do not do anything to salvage them.

Then, in "Why Libet-Style Experiments Cannot Refute All Forms of Libertarianism," László Bernáth argues that such experiments are able to serve as evidence against forms of libertarianism that do not make metaphysical distinctions between types of decisions. However, he claims that there are a class of libertarian positions for which they are powerless: those that restrict the set of free decisions in a way that rules out their testing in existing paradigms (though he suggests ways in which those paradigms might be modified).

In "Actions and Intentions," Sofia Bonicalzi argues that recent findings in cognitive neuroscience militate against a proposition-style causal theory of action. Instead, she claims, we are better off thinking of action as the product of complex interactions between a number of different systems. Under such a scheme, intentions are not plausibly context-independent, inherently causal, discrete entities. On the basis of specific Libet's experiment interpretations, she suggests that neuroscience can play a constructive role with respect to basic concepts in the philosophy of action.

Finally, the third part of the book returns to the articulation of language and causality in agency and free will in human beings. Anna Drozdzewska argues that the problem of mental causation is of central importance in the free will debate, despite the fact that it is often missing from discussions in the extant literature. In "The Mental, the Physical, and the Informational" she suggests that the right approach in this context will be to consider the causal role that information can play in the brain. She motivates the view that this may provide a new approach to the problem of causal exclusion.

Last, in "Free Will, Language, and the Causal Exclusion Problem," Bernard Feltz and Olivier Sartenaer address a similar theme, by considering the ways in which the use of language might instantiate emergent causal powers that produce downward-causal effects. In doing so they bring recent ideas about

diachronic causation into contact with the neuroscience of learning and philosophy of language.

4 Opening Perspectives

The question of free will admits of a diverse number of positions. Even in this book, which brings together contributions by authors largely open to the possibility of free will, it would be hazardous to propose general conclusions to which all could agree. However, in our role as editors of this book, we would like nonetheless to propose some final thoughts on the relation between language and causality with respect to the question of free will.

From a social point of view, it seems difficult to defend the idea that language lacks causal power. We need just be reminded that science is language, and that the products of science—technology—are unthinkable without the causal efficacy of language. Law, economics, political science, rhetoric, all of these are equally languages for which it seems superfluous to argue for their efficacy. Now, if language has causal power from this social point of view, the monist presupposition requires that we posit its efficacy at the individual level as well. If one wonders about the causal efficacy of language at the level of the individual, then in a certain sense the question is *how* to understand this efficacy and not *if* there is any such efficacy. Language operates just as much on the individual level as it does on the social.

On this point, the contributions on language in this collection demonstrate that it is possible to think about the effect of language on the brain while respecting the principle of causal closure. Such a result is important since it allows us to give language a decisive place in our thinking about free will and to bring about a rapprochement between certain philosophers of language and the experimental work of contemporary neuroscientists. In decision-making processes, language is not epiphenomenal. It is perfectly coherent to defend that language has a causal influence in decision-making processes.

However, this causal efficacy of language does not correspond to free will. For example, some thinkers inspired by structuralism defended the idea that language itself determines behaviour through the unconscious (in Lacan's (1966) "Return to Freud") or determines collective behaviour through ideology (as in Althusserian (1970) Marxism), and without the knowledge of the persons concerned. So in order to intervene in the debate over freedom on the basis of language, one needs to go beyond its efficacy.

In contrast with structuralism, Habermas (2007) develops the idea of a productive language in culture that gives rise to meaning: this is what he calls

"objective mind." The human subject, by learning language, becomes part of this cultural dynamics and becomes able to participate in social conversation, becomes capable of inventiveness, creation, and novelty. The latter is what Habermas calls "subjective mind." For Habermas, then, language is not an oppressive structure but an open and dynamic structure that allows each individual within it to choose the meaning of her existence. To be free is to behave according to the system of meanings that we have chosen and helped to construct. Not only is language causally efficacious then, but it brings with it the potential for novelty.

Habermas's philosophy of language leads us to defend the idea that the causal power of language allows for a wider efficacy of the processes that involve the self-determination of human behaviour. Language, through which the individual participates in objective mind, allows each person to construct a system of meaning by which he gives sense to the world and orients himself behaviourally. Neural plasticity and learning as implemented in neural networks are the mechanisms which allow to understand the effect of language on the brain. These mechanisms bring us back to these processes themselves as the conditions of possibility for the production of language. Such a perspective makes it possible to meet the difficulties related to the problem of causal exclusion. It produces reconciliation between neuroscience and a conception of the human being as free.

References

Althusser, L. (1970). Idéologie et appareils idéologiques d'État (notes pour une recherche). *La Pensée*, 151, 3–38.

Atlan, H. (2011). *Selected Writings: On Self-Organisation, Philosophy, Bioethics, and Judaism*. New York: Fordham University Press.

Badler J, Lefèvre P, Missal M. (2010). Causality attribution biases oculomotor responses. *J Neurosci.* 2010 Aug 4;30(31):10517–25.

Badler JB, Lefèvre P, Missal M. (2012). Divergence between oculomotor and perceptual causality. *J Vis.* 2012 Jan 1;12(5):3. DOI: 10.1167/12.5.3. Print 2012.

Born, M. (1949) *Natural philosophy of cause and chance*. Oxford: Clarendon.

Craver, C.F. (2007). *Explaining the Brain. Mechanism and the mosaic unity of neuroscience*. Clarendon Press, Oxford.

Glymour, C. (2001), *The Mind's Arrows : Bayes nets and causal graphical models in psychology. Cambridge*, MA : MIT Press.

Habermas, J. (2007). The language game of responsible agency and the problem of free will: how can epistemic dualism be reconciled with ontological monism? *Philosophical Explorations*, 10, 13–50.

Hume, D. (1739/1987). *A Treatise of Human Nature*, Ed 2 (Mossner EC, ed). Oxford: Clarendon.

Lacan, J. (1966). *Écrits*. Paris: Seuil.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioural and Brain Sciences*, 8, 529–566.

Michotte, A. (1949/1963). *The Perception of Causality*, (Miles TR, Miles E, transl). New York: Basic Books.

Pacherie, E. (2008). The Phenomenology of action: a conceptual framework. *Cognition,* 107, 179–217.

Russell, B. (1913). On the Notion of Cause, *Proceedings of the Aristotelian Society,* Volume 13, Issue 1, 1 June 1913, 1–26.

Scholl, B.J., Tremoulet, P.D. (2000). Perceptual causality and animacy. *Trends Cogn Sci* 4:299–309.

Wagemans, J., van Lier R., Scholl B.J. (2006). Introduction to Michotte's Heritage in Perception and Cognition Research. *Acta Psychol* (Amst) 23:1–19.

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

Woodward, J. (2003). *Making Things Happen*, Oxford University Press.

PART 1

Intention and Consciousness

CHAPTER 1

Perceptual Decision-Making and Beyond: Intention as Mental Imagery

Andrew Sims and Marcus Missal

The standard view in the philosophy of action is the Causal Theory of Action (CTA).[1] On this view, a behavioural item counts as an intentional action if and only if it is caused by the appropriate sorts of mental states in the right kind of way. On most popular contemporary accounts (e.g. Searle 1983; Bratman 1987; Mele 1992; Pacherie 2008), the appropriate sort of mental state is an *intention*, with the intention construed as an attitude towards a proposition. The "right kind of way" is thought to be one in which the content of the intention propagates from an abstract level of description (e.g. the intention to investigate a noise) through to more fine-grained specifications that give a bodily movement a rational structure at the moment of bodily movement (e.g. the intention to switch on this light), and finally culminating in motor commands required to execute the right bodily movements. These levels of abstraction respectively correspond to so-called distal intention, proximal intention, and motor intention. These three kinds of intention have distinct roles in the overall dynamics of intentional action (Pacherie 2008; Mele this volume).

In our contribution to this volume we offer an alternative theory of intention, on which it is not a propositional attitude at all but rather a distinct kind of mental imagery. For the purpose of our argument, we can provisionally define a mental image as a quasi-perceptual representation that occurs in the absence of the corresponding stimuli. Such imagery need not be conscious; it can also be unconscious. It may manifest in one or more perceptual modalities (Nanay 2017). The main difference that we wish to mark is that mental imagery has a quasi-perceptual format rather than a quasi-linguistic or propositional one. That idea will be developed in more detail in Sections 1 and 2.

Our account is inspired by work in the perceptual decision making literature, where decision is modelled as a process of evidence accumulation under conditions of uncertainty and noise. In the paradigms that are central to this

1 This work is supported by the Action Recherche Concerté project "Causality and Free Will." We would like to thank Anna Drozdzewska and Daniel Burnston for some helpful comments on an earlier draft.

literature, the subject is required to make a simple movement (e.g. look left or right) depending on a perceptual condition (e.g. a visual signal which indicates the "correct" direction). As time passes, the accumulation of perceptual evidence biases the drift of a single value called a decision variable, which when it hits a bound (i.e., becomes equal to a specified value, or set of values, if the variable is multidimensional) propagates a signal to the relevant effectors and causes the bodily movement. Models of this kind have great predictive power when it comes to the way that error rates ("wrong" decisions) and response times are distributed in experimental samples.

It could be argued that these studies lack ecological verisimilitude: not many decisions are of this nature (though some are, like crossing the street when the light turns green). However, it's nonetheless possible that the model we describe could profitably be extended to contexts beyond perceptual decision making. For example, it's been suggested that in cases where an action is more or less endogenous (i.e. not simply produced in response to a cue), the same model could be used without there being any specific perceptual evidence to be integrated with the decision variable (Schurger et al. 2012; Murakami 2014; Schurger et al. 2016). In this case, Schurger et al. (2016, 78) argue, "the process of integration to bound is dominated by ongoing stochastic fluctuations in neural activity that influence the precise moment at which the decision threshold is reached." In other words, in these cases it is just neuronal noise which biases most strongly the decision variable, without any real perceptual evidence to speak of.

In this context, we should certainly want to know what it means for something to be intentional: if the evolution of the decision variable during spontaneous voluntary movement is dominated by stochastic drift, then this would appear to make the outcomes of our decisions implausibly random.[2] With this in mind, our suggestion is that intentions can be understood as mental images that play the role usually reserved in this model for bona fide perceptual evidence (e.g. the onset of a visual cue). Intentional action is therefore to be

2 We do not mean to attribute such a view to Schurger and colleagues. Those authors do suggest that there can be biasing influence from "anticipation, subjective value [and] clock monitoring" and that "parameters, such as the movement type, may be fixed by a prior decision, which may in turn channel the neural activity leading up to the decision threshold towards a specific effector." (Schurger et al. 2016, 78) Presumably what they have in mind when discussing neuronal noise is endogenous voluntary movements where the agent has no strong commitment to any outcome, like in Libet-style experiments. But proposals beyond these are largely unspecific at present; it's our ambition to put some meat on these bones in a way that addresses extant problems in the philosophy of action. We are much inspired in this by Burnston's (2017) programmatic remarks.

construed in terms of this evidence-accumulation mechanism, but with its scope expanded so that inner states as well as evidence can play a biasing role.

We see the following merits in this account. First, it sidesteps some pressing problems in the philosophy of action; these are problems that we suspect result from the theory of intentions as propositional attitudes. For example, there is an issue concerning the interface between propositionally-formatted representations and the representational format that is appropriate to fine motor control. That is to say that there is confusion in the literature over how the content of a proximal intention could propagate to a motor intention (Butterfill & Sinigaglia 2014; Mylopoulos & Pacherie 2017). Indeed, some authors express scepticism that this is possible at all (Burnston 2017). On our account, there is no need to explain such an interface because the representational format that is appropriate to sensorimotor control is common to *all* levels of intention.

Second, it expands on the Schurger et al. (2012) interpretation of Libet-style experiments and the sceptical arguments that issue from them. The orthodox interpretation is that the readiness potential (RP) represents the final stages of action planning and preparation, and that this occurs *prior* to the conscious intention to move. That is why it is used in an argument that decisions are never caused by conscious intentions: the putative decision (the RP) occurs *before* the conscious intention. Schurger and his colleagues argue convincingly that this interpretation is wrong: we should interpret the RP as the stochastic drift of the decision variable, which ramps up to hit the bound at about the same time as the agent feels the "conscious urge to move." Our account will offer a framework for extending this proposal to action more generally, and for the non-sceptical interpretation of Libet-style paradigms and others like them.

However, to explicate it persuasively will be challenging. There will be three philosophical objections to our thesis that we must address. First, the mechanism we describe entails that action-producing mental imagery stands in competition with bona fide perceptual evidence issuing from the senses. That obliges us to explain how action ever gets off the ground: how is it that motor imagery detailing the aimed for state of affairs can bias the decision variable in a way that moves it away from what is currently perceived?

Second, it is unclear how mental images could play the quasi-inferential role in the rational control of action that propositional attitudes plausibly can. Practical reasoning can be expressed in a deductive form quite easily if the states involved are propositional, since their syntactic structure allows this. In this way, we can understand how standing intentions could constrain practical reasoning and the implementation of other intentions. A corresponding account needs to be given for mental imagery.

Finally, we need to be able to show that perceptual or quasi-perceptual content can represent a state of affairs at the level of abstraction that is appropriate to propositions, and without themselves being propositional. For instance, when I intend to go on a holiday at some point in the future, but without any specific place in mind, I have a distal intention which specifies an action at a very high level of abstraction. But we usually think of mental imagery (like perception) as being determinate with respect to all its details, and not abstract in such a way.

These important challenges for our account will be met in the course of the paper. But we must begin by discussing the hypothesis in more detail.

1 Perceptual Decision Making and Diffusion-to-Bound Models

A neural decision variable could be defined as a compound signal containing sensory evidence, priors, value, and other quantities needed to make a categorical choice. This choice results in a rule applied to the decision variable i.e. if enough evidence is accumulated in a certain context, I will choose to do *this* rather than *that*. Decisions based on perceptual evidence are often conceptualized using diffusion-to-bound models. In these models, a hypothetical neural decision signal accumulates until it reaches a threshold value (i.e., the bound) causing movement initiation. Response time and accuracy (or the percentage of correct responses) are modelled using a noisy evidence accumulation process like a continuous random walk or diffusion (Ratcliff & Hacker 1981). The diffusion process encodes information about the experienced stimulus and other sources of information present in the context that are needed to make a decision (Figure 1.1).

This theory was initially developed to model experimental results in a sequential letter-matching task where human subjects have to judge whether two strings of letters are identical. Typically, the two strings of letters are rapidly presented sequentially on a computer screen facing the subject. The subject has to compare the previously encoded string of letters held in working memory with the last one. The decision the subject has to make is: "Are the two strings the same or different?" Technically, this type of task is often referred to as "two-choice serial reaction time task" (2-CSRTT).

According to Ratcliff and Hacker (1981), if the diffusion process reached the upper bound, a "same" response was produced. Alternatively, if the diffusion process hit the lower bound a "different" response was produced. The diffusion rate towards the upper bound increases as a function of the proximity of the two strings and increases towards the lower bound as a function of the

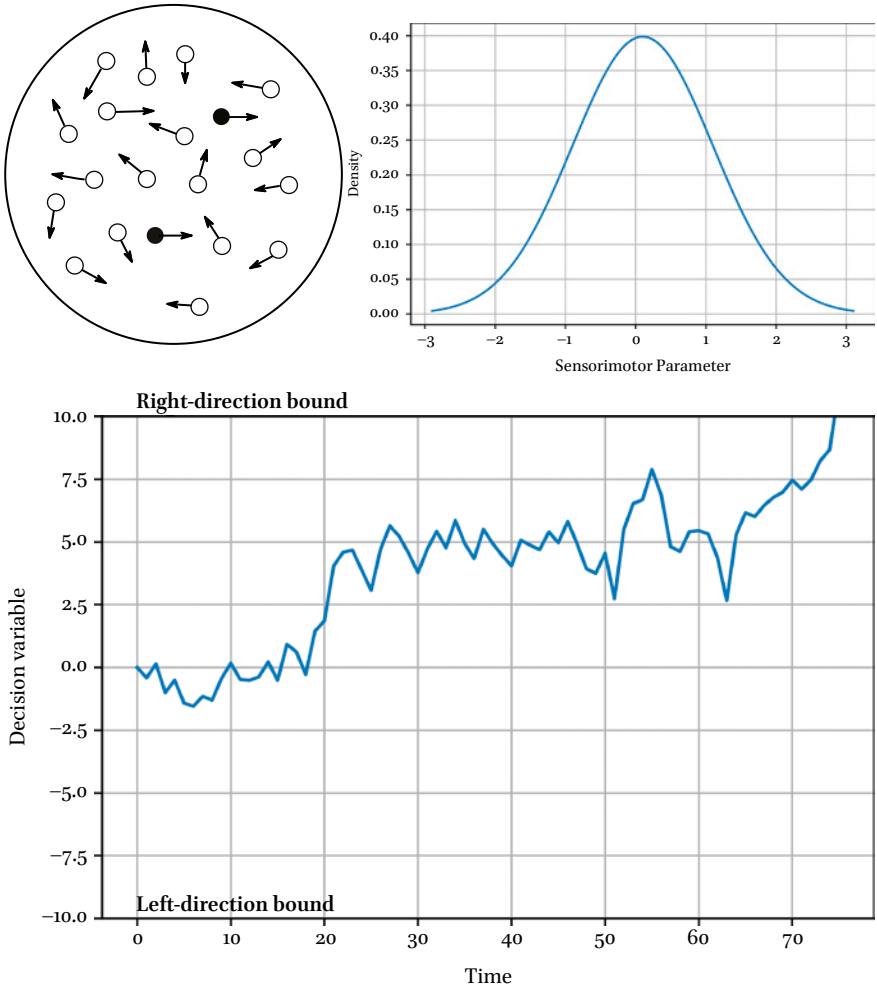

FIGURE 1.1 A (top-left): This is an illustration of the visual stimulus that is presented to the subject during the dot motion paradigm. During a single trial the dots all move in random oblique direction except for a specified proportion which move either to the left or to the right. In the trial shown, 10% of the dots (those which are darkened) are moving to the right; B (top-right): This shows the evidence that the visual stimulus of Fig. 1.1A is taken to comprise. It is a normal distribution with mean = 0.1 and standard deviation = 1. The sensorimotor parameter is the direction of movement of the dots in the visual field. In this simulated and merely demonstrative example, the relationship between the stimulus and the distribution should not be taken as necessarily accurate; C (bottom): This shows the activity of the diffusion-to-bound mechanism over a single trial. At each time-step, a random value is sampled from the distribution in Fig. 1.1B and added to the decision variable. This continues until the variable hits the bound corresponding either to left-hand movement or right-hand movement. In this case, the decision variable hits the move-right bound after about 80 time-steps.

dissimilarity of the two strings. Moreover, human subjects probably implicitly adjusted the position of the boundaries and the zero point of the diffusion process (bias) based on previous experience or expectation. For instance, giving an instruction to the subject before data collection (e.g. answer "same" only if you are sure *versus* answer "different" only if you are sure) changes the bias in the model and therefore reaction time is prolonged but accuracy is increased.

Several experiments in the Rhesus monkey have tried to find areas in the brain where the accumulation of evidence and/or the decision variable could be processed. The major problem is to differentiate accumulation of sensory evidence from a decision signal at the single neurone level. Indeed, the two signals are inevitably linked. Therefore, perceptual decision making paradigms have been developed in order to separate in time these two different signals (see the review in Gold and Shadlen 2007). For instance, in the random-dot motion paradigm, Rhesus monkeys or human observers have to decide whether a set of random dots move to the left or to the right. The percentage of coherently moving dots is randomly and parametrically varied between successive presentations (Gold, J.I., Shadlen, M.N. 2002). Even at zero coherence (all dots moving randomly) subjects must make a decision.

Shadlen and co-workers have shown that a neural correlate of the decision variable involved in this process can be found in the parietal cortex (area LIP; Shadlen et al. 2006; Roitman & Shadlen 2002). Interestingly, neuronal activity also reflects the choice of subjects at zero coherence (no sensory evidence per se). Activity in area LIP is compatible with predictions of the diffusion-to-bound model. Lafuente & Romo (2005) have reached similar conclusions using a task in the somatosensory domain where Rhesus monkeys have to detect a 20-Hz sinusoidal indentation applied on the finger pad. This simple detection task was accompanied by neuronal recording either in the primary somatosensory cortex (S1) or the medial premotor cortex (MPC). S1 neurons encode the presence of the sensory stimulus, but not the decision variable. On the other hand, MPC neurons are good candidates to encode the decision of the monkey in the presence (or absence) of the vibrotactile stimulus.

All these experimental paradigms provide strong electrophysiological support for the diffusion-to-bound theory of decision making. However, most real life decisions involve deliberation over topics where repeated experience is usually not available (e.g., a first marriage). Perceptual decision experiments are often rather simple, linking one particular sensory quantity to one movement (e.g. a button press, an eye movement). It is therefore tempting to posit different mechanisms for decision making, one for simple sensorimotor decisions and one for more complex deliberate decisions (Kahneman 2002).

However, we suggest that it is more plausible to postulate one decision making mechanism that could be instantiated in different areas depending on the context. Indeed, in humans even simple decisions are actually often influenced by deliberation. The aim of the present paper is to speculatively extrapolate from the theoretical contribution of diffusion-to-bound models to the kinds of decisions discussed by philosophers of action.

2 Extending the Model

As noted, perceptual decision-making paradigms do not accurately represent all decisions made by human beings in the wild (i.e. outside the laboratory), except perhaps in a highly idealised way. That is because human beings in the wild also perform actions more or less endogenously, that is to say, without any obvious perceptual cue that directly triggers the decision. Such choices usually also take place over more than two alternatives. However, the models for perceptual decision making can be extended to account for these more complex cases. Such extension is relatively speculative but it is not without support.

In the paradigms discussed above, there are usually just two alternatives. For example, in the dot movement paradigm the choice is whether to indicate that the dots are moving left, or whether to indicate that they are moving right. One of these choices will be incorrect, depending on the perceptual cue (e.g. if I indicate left when the dots are moving right). Simple choices like these are easy to model: as we have shown, this can be done in two dimensional space where one decision corresponds to time and the other corresponds to the value of the decision variable. In such a two dimensional model we can specify two bounds that correspond to two outcomes, but we cannot add any more.

Therefore, to model more complex decisions requires that we add more dimensions to our model. That allows us to add more bounds in order to represent more possible outcomes. For example, we might model a decision variable that drifts stochastically through three dimensional space. Strictly speaking, the decision variable will have two dimensions <x, y> while the third dimension <z> represents the evolution of the decision variable through time. In this three dimensional space we can add more reachable bounds that represent more possible outcomes. The bound will be defined by a hyperplane, which in a three dimensional space is just a subspace with two dimensions. At each step in the drift of the decision variable, two values are added (plus the time step) that dictate how far it moves in this space. Theoretically speaking there is no limit to the number of dimensions we can add to our model, despite the fact

that it becomes difficult to visualise after the fourth is added. So in principle, we can represent any number of possible outcomes within this diffusion-to-bound style of model.

The dimensions of our model correspond to sensorimotor parameters that vary continuously along a scale. Examples of these parameters from vision may be the size of an object, or its location in the visual field. Similarly, these parameters may encode for motor features that have continuous properties, like extension of reach or strength of grasp. For a particular bodily movement, then, the dimensions which are integrated into the decision variable at each moment reflect the sensorimotor parameters for which these dimensions encode. This allows us to improve on our provisional definition of "mental image." In the context of this paper, we propose to understand a mental image as a set of values in n dimensions which correspond to a range of sensorimotor parameters, where this range is minimally limited to a specific value in each dimension that the image includes, but which may also specify a wider range. So, for example, a mental image may specify a set of highly constrained values that correspond to my extending a finger to touch a specific point in front of me, but it may also specify a looser range of values that correspond to my having an arm extended out in front of me, without any particular point being specified.

Hommel's (2004) theory of event files suggests a specific way in which we might construe the role of sensorimotor evidence during endogenous action. Event files are episodic bindings of perceptual features, a task context, and motor features. Event files specify a particular directed bodily movement. For example, it might specify the action of reaching for a bottle of water. In this case, the event file would bind the perceptual features corresponding to the action (proprioceptive, visual, and other sensory items associated with successful reaching) with motor commands that correspond to the correct extension, rotation, and grasp.

In the context of diffusion-to-bound models, it may be that we can understand the bounds as corresponding to event files, the sense that the event files specify the appropriate locations in state space for the execution of a particular action. Then, different influences bias a decision variable in the state space by accumulating evidence for one or another event file. The evolution of the decision variable through the state space represents the competition between event files for implementation in behaviour (cf. Cisek & Kalaska 2010). The outcome of that competition is jointly determined by stochastic drift (i.e. neuronal noise) and biasing influences. These influences are understood here to include the following two sorts of item: i) sensorimotor values that are caused by stimuli (i.e. perception per se); and ii) sensorimotor values that are not

caused by stimuli (i.e. mental images). These two sorts of items both serve as biasing influences that drive the decision variable to a bound.

Here is a simple case which is intended to make the proposal clearer. You are standing in front of a door. Then you open the door and walk through. What happened? A certain kind of CTA-theorist will say that you had an intention with propositional content corresponding to something like "open the door and walk through." This intention propagated from higher cognitive areas into motor cortex and was somehow translated into the appropriate motor commands, and then bodily movements.

Our story is different. The right bodily movements correspond to a set of sensorimotor values that are executed in the right sequence. The decision variable does not correspond to these parameters before you act; it is rather fluctuating around the values associated with the current perceptual state (the closed door). Action begins when biasing influences begin to push the variable towards the values associated with the successful execution of the action. The influence in question will be a set of endogenously generated quasi-perceptual representations—that is, a mental image—that is implemented in probability distributions over the appropriate values in those sensorimotor parameters. The decision variable is continuously updated with randomly sampled values from these distributions, and this drives the value of the decision variable to the event file that corresponds to you having opened the door.

A story like this can be extended upwards to more paradigmatically intentional action and it can also be extended downwards to what philosophers call "sub-intentional" action (O'Shaughnessy 1980). Sub-intentional action is purposive activity that is not intentional under a description, like idly doodling on a notepad while focused on something else, or tapping impatiently while waiting for something. Actions of these kinds can be construed in our model as biasing influence that results from learned association of perceptual values with corresponding "stereotyped" motor values. For example, it may be that a pen lying next to a pad is associated with the action of doodling, and so automatically potentiates the sensorimotor parameters associated with doodling. This is what should normally happen in the absence of motor inhibition (cf. Duque et al. 2017). In cases where we are focused on something else, we fail to inhibit those actions that are potentiated. Furthermore, the capacity for such motor inhibition may be reduced by brain lesion, in cases such as in utilisation behaviour where the affected part of the body engages with nearby objects in a way that is contextually inappropriate (Shallice et al. 1989).

The upward extension is more complex. It requires that endogenous sensorimotor variables can bias the evolution of the decision variable at great temporal and spatial distances from the corresponding action. For example, if I want

to walk to the cafeteria on the other side of campus to get myself a coffee, then this requires that I am able to potentiate the right sensorimotor parameters in the right sequence, and in a way that respects contextual demands (e.g. not taking a route that I know to be closed). We think that in these more complex cases propositional attitudes have a role to play, but it is not the role that is envisioned by CTA-theorists. We propose to elucidate this upward extension of our account by addressing a few key objections from the perspective of CTA.

3 Some Conceptual Problems with the Account

First, it's not clear why mental imagery would sometimes win competition with percepts that are immediate to the agent. Secondly, we might doubt that mental imagery as a representational format is of sufficient conceptual complexity to be able to exercise the rational control that is required of an intention. Finally, despite its evidential support in simple contexts, we might doubt that models for perceptual decision making could be expanded to contexts in which we intend to perform an action that is specified at a very high level of abstraction. In this section of the chapter we address these concerns.

3.1 *Can Mental Images Produce Action?*

A critic of our view might deny that mental imagery can drive action, which would make it unsuitable for playing the role of intention. The reasons that might be cited for this view are as follows. First, there is the commonsense idea of mental imagery and imagination in general as a kind of derivative version of ordinary perception. The similarity between imagery and perception is noticed by many authors in the history of philosophy, most notably by Hume (1739) in his claim that impressions and ideas are distinct only in their "force and vivacity," and indeed something like this comparison motivates representational views of perception in which the representational content is made to do work in distinguishing hallucination from veridical perception.

Secondly, it might be argued that the default attitude to hold upon the apprehension of some content in imagination—insofar as any attitude is taken up at all—would be belief, rather than intention or desire. That is to say, if I imagine something, and (in some exceptional case) I forget or I am unable to know that I am merely imaging, then surely it is more correct to say that I *believe* my eyes, rather than that I *intend* my eyes. So here, the objection will be to say that for the mental image to play the role of an intention requires something more than the image. And presumably the suggestion will be that the additional ingredient is some sort of conative intention-like attitude towards that image, which would make our account redundant.

However, these objections and others like them rely on a commonsense notion of mental imagery or imagination which we are not committed to. On the mechanism we have hypothesised in Section 2, what is given as input to the mechanism is just a set of sensorimotor parameters which correspond to the consequences of the action, and which drive it to a bound. This means that there is no real representational distinction of type between a perceptual representation and the endogenous representation ("intention") that produces action. This feature of our approach—despite how odd it may seem to philosophers—is not unprecedented. Prinz (1997), for example, outlines the "common coding hypothesis," on which perception and action planning share a representational form in common. That is to say that items that are "efferent" (i.e., motor) and those that are "afferent" (i.e., sensory) are continuous and overlap; this would mean that none of these representations code specifically for afference or for efference, or if they do, then they code for both simultaneously, as in Millikan's (1995) hypothesised "pushmi-pullyu" representations.[3]

There are empirical reasons to think that this common coding hypothesis is true. If "event codes" and "action codes" encoded respectively for afference and efference, then we would expect their respective functional roles to be well defined. Accordingly, we would not expect there to be interference or facilitation effects on action by way of perception, or vice versa. The interference hypothesis states that when a particular code (e.g., corresponding to location) is in use for perception, it will be impaired in its use for action (and vice-versa). The facilitation hypothesis states that dimensional overlap between a stimulus and a response can enhance the speed of the response. These sorts of effects are often reported in the literature (see Prinz 1997; Hommel 2004 for reviews; cf. Burnston 2017), which motivates for the view that direction of fit is not part of these representations.

All this, of course, obliges us to answer the following question: when, and in virtue of what cause, is it endogenous mental imagery rather than perception which brings about action? That we have a mental image in working memory must certainly not be sufficient for bringing about action; sooner or later we must accept that such images simply do not correspond with the deliverances of perception. So how do such images can drive action, in the absence of an encoded direction of fit that allows it to be distinguished from perception? Someone taking our view seriously might think that veridical perception prevents action ever getting started, because this constant source of evidence will bias the decision variable in preference to any endogenous imagery.

3 This corresponds to the philosophical distinction between "directions of fit" (see Anscombe 1957, §32 for a classic exposition).

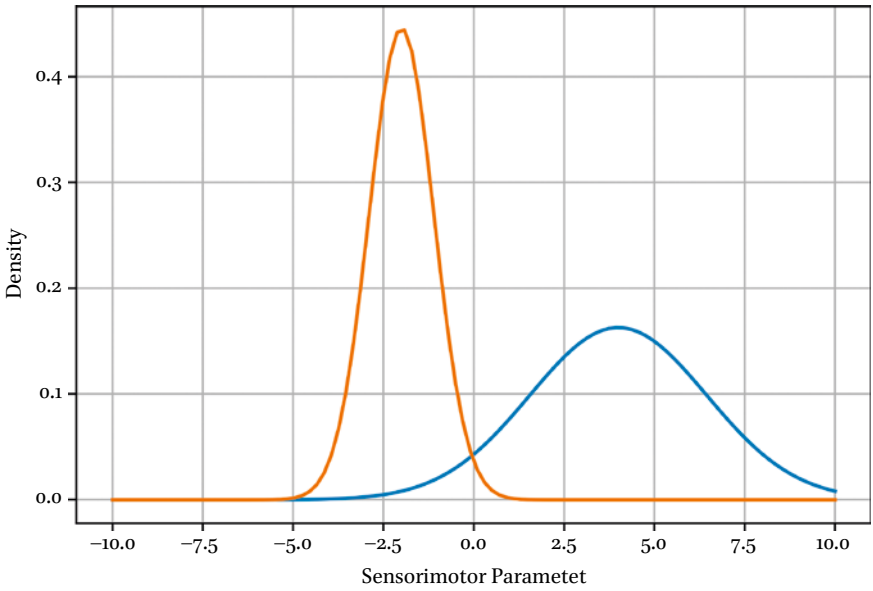

FIGURE 1.2 This is a demonstration of the concept of precision with reference to two sensorimotor parameters. The distribution on the left has a much lower precision. As such, a sample from it may fall within a much wider range of values. The distribution on the right has a much higher precision. Our claim is that when values from distributions are sampled for a mechanism of the kind illustrated in Figure 1.1, the decision variable will reliably tend towards the values specified by a more precise distribution.

We can develop the mechanism of Section 2 in order to answer this question.[4] The inputs to the decision variable at some time consist of an array of values that correspond to sensorimotor parameters. These values, however, are not discrete: given the conditions of uncertainty and noise that prevail, these values are more likely to be probability distributions. Probability distributions have various properties. One of these properties is *precision*. The precision of a probability distribution is inversely related to its standard deviation; it is a measure of uncertainty. A distribution which represents a perceptual parameter with a very high level of uncertainty will have a low precision; values that are sampled from that distribution will not cluster very tightly around the mean. The converse is true of distributions with a high precision (Figure 1.2).

With this in mind, our suggestion is that whether or not some mental image is perceptual or action-driving is a function of the precision of the corresponding

4 We are inspired by similar ideas in the context of predictive processing (e.g., Hohwy 2013; Clark 2016). However, nothing we say here commits us to that framework as a whole.

distributions. Specifically, mental imagery will produce action, rather than be revised on the basis of sensory information, in cases where it is higher in precision than the sensory information that is in tension with it. When it has a high precision the values that are integrated into the decision variable will be closer to the goal-state value than if the distribution is imprecise. Similarly, when the current perceptual state is relatively imprecise then the values that are sampled from that state will be relatively ineffectual in preventing the biased diffusion of the decision variable to the goal state.

To conclude this section of our argument, we accept that our initial account was problematic in the context of philosophical orthodoxy regarding direction of fit. However, we believe that this distinction is only important in the present context because it explains why a content would sometimes be afferent (responsive to informational change) and sometimes be efferent (action-producing). In response to this concern, we have suggested how we can explain this in the context of an account that does not rely on the idea that there is an action/perception distinction built into the representations that underlie cognitive processes.

3.2 *Content and Rational Control*

The second conceptual problem that we wish to address corresponds to the role of intentions in allowing rational control over action as it unfolds in a particular situation. This control has a number of different aspects. Pacherie (2008), following Buekens et al. (2001), classifies these into two. The first is *tracking control*. Tracking control is our ability to rationally control an action as it unfolds in time. For some action, it will include various steps whose temporal sequence must be respected, and it may be that we are required to implement an intention in an unforeseen manner if we are thwarted at any of these steps. Take the action of fetching a cup of coffee. To successfully perform this action I have to leave my office, get to the kitchen, and set up the coffee machine to make a cup. Tracking control of this action will involve performing this action in the right sequence (e.g., not trying to extract the coffee before the ground beans are in place) and improvising on the fly (e.g., if there is work being done in the corridor, I will have to take a detour to the kitchen from my office).

The second is *collateral control*. This involves controlling for side effects of my action. When I go to fetch my cup of coffee I may wish to perform the action in a way that respects other commitments that are in the background. For example, it may be that I implement the intention by taking a circuitous route in order to avoid encountering someone I don't wish to see. I may also wish to implement my intention in a way that respects etiquette: if there is someone

already waiting for the coffee to brew, then I will not cut in line to take my coffee before they do. The bottom line for both of these types of control is that they need to be responsive to, i) other commitments I have, even if dispositional (the commitment to acting politely); and ii) the exigencies of the present context in the here-and-now.

For certain instances corresponding to both of these sorts of control, it may be argued that the intention needs to be poised to play a role in practical reasoning. If I find myself thwarted by contextual constraints (work in the corridor), then I will need to engage in practical reasoning on the basis of the intention in order to revise my plan of how best to implement it. For collateral control, my intention needs to be rationally responsive to more abstract dispositional states regarding certain constraints within which I would like to bind my behaviour. So it would appear to be the case that both of these types of control involve inferential relations between the intentions and propositional attitudes like beliefs and desires.

Given the requisite inferential sophistication of intentions, one criticism of our proposal might be that mental images are inapt to play that role. That is because inference seems to be something that is appropriately performed in an amodal conceptual format, rather than a quasi-perceptual format. It is hard to see how operations like material inference, disjunction, and so on could be implemented in the latter, but these are arguably required in order to exercise rational control over action. To see this, say that I have the intention to get a cup of coffee and there are two different paths that I could take to the kitchen. I also want to pass by my colleague's office on the way, and this is on one of the paths. So I somehow have to be able to implement the intention in a way which respects this other intention, and that requires that my intention is rationally responsive to the other. If intentions are conceptual or language-like, then it is easy to see how this would work: just in the same way that inference is performed in written or spoken language. But this is not possible on our account.

Our reply to this objection is to advance a clarification. It is not a consequence of our position that propositional attitudes can play no role in action production whatsoever; our contention is just that they do not play any causal role *qua* intentions. It may well be the case that propositional attitudes play a key role in relations between images, and this would not necessitate that intentions are propositional attitudes. For example, van Leeuwen (2013) argues that propositional attitudes like beliefs can inferentially govern transitions between imaginings. Here's how he demonstrates the point:

> Instructions: imagine a female lion in the bushes looking out on a herd of grazing antelope; then imagine she charges them. What do you imagine

> next? I'm guessing you imagine the antelope run away. But why? Imagining this wasn't in the instructions. You could have imagined them staying put, though you didn't. The answer, around which some consensus has emerged, is that one's *beliefs inferentially govern transitions from initial imaginings to later imaginings*.
>
> VAN LEEUWEN 2013, 226

We can imagine a wide range of examples of this basic type, where the way mental imagery unfolds in time is governed by our dispositional mental states. I can "deliberate" in imagination when it comes to settling on an intention (Nanay 2016), and my deliberations will be constrained by what I believe it is possible for me to do, and what I think is the likely result of some imagined action. Furthermore, we see no reason to limit this governing role to beliefs. It seems that desires may play a key role in imaginary transitions, as well. Certainly, this role is familiar enough from the quite widespread and ordinary experience of pleasant daydreams. In daydreams, transitions between imaginings are often governed by what we would like to happen, and also by our beliefs about potential ways in which such things might occur.

The constraints that propositional attitudes place on transitions in imagery provide us with a way to understand how mental images *qua* intentions could exercise the kinds of control discussed above: tracking control and collateral control. Let's take the previous case of going to get a cup of coffee. As I implement the intention, it's necessary that I implement it in a way that respects a certain sort of temporal sequence. If I try to extract the coffee before I load the ground beans into the machine, then my action will not succeed. This is rational control as tracking control. Mental images can exert this kind of control since transitions between images is governed by our beliefs about the world. I have certain beliefs about the way this coffee machine works, and these beliefs govern the way in which my imagery-intention represents the successive stages of the action that I am engaged in.

Likewise, when I intend to get my cup of coffee, I need to implement my intention in a way that respects my other commitments. If I go and get it in a way that violates these commitments, then I will not be exercising collateral control over my action. Collateral control can be explained by the way that our desires govern transitions between mental images. For example, I may have the background commitment to talk to a colleague on my way to the kitchen. This is to say that I desire that I talk to my colleague. This desire will govern the way my intention-imagery represents successive stages on an action, by governing the transitions between image stages. So the intention will be that I go by the circuitous route to the kitchen on which my colleague's office is located.

In short, there is no reason to think that images cannot exert the right kinds of rational control over action once the role of bona fide propositional attitudes is taken into account. Our claim is just that propositional attitudes do not themselves play the role of intentions, as the prevailing orthodoxy states. Why then, would this not make our account of intentions as mental images superfluous? In other words, aren't we back now at an orthodox view that includes *both* mental images and propositional attitudes? That is not the case; our position differs in a subtle but important respect. Specifically, on our account there is no propagation of content from propositional attitudes to mental images (i.e. motor representations), whereas that is a central feature of CTA (cf. Pacherie 2008). For us, propositional attitudes merely play an auxiliary role in constraining possible transitions between mental images.

The reason that this difference is important is because a key puzzle that drives contemporary philosophy of action—the "interface problem"—concerns how it is that content could propagate from propositional representations to sensorimotor representations. Pacherie's framework is subject to this problem, because for her there is such a propagation. But for us there is not, and we are therefore providing a way of dispensing with this vexing philosophical problem about content.

A critic might still object that we are committed to some propagation of content as soon as we accept the view that propositional attitudes could govern transitions between images. That is because such governance requires that the content of the proposition plays some sort of causal role in the transition, and this amounts to the propagation we claim to be entitled to deny. To this objection we can give a suggestion about how this might occur in the absence of direct translation from propositional content to sensorimotor content.

Let's return to Van Leeuwen's example of the lion and gazelles. In this case, there is a transition between the image of the lion charging these gazelles to an image of the gazelles fleeing the lion. Let's say that this transition is governed by a belief that gazelles (generally) flee from lions. Our critic will say that what happens here *must* be some sort of propagation of content between that relevant proposition and the image, because the proposition seems to have causal influence over the evolution of the episode of imagining. And that would oblige us to give an answer to the interface problem.

However, we don't see that mere causal influence amounts to content-propagation. For example, it may be that the proposition <gazelles flee from lions> activates a number of high level sensorimotor parameters without there being any direct translation of the propositional content into sensorimotor imagery. For example, it may be that due to associative processes the tokening of the lexical item which expresses that proposition *biases* the decision variable

towards the range of sensorimotor parameters which correspond to a statistical regularity that holds between its contents (c.f. Burnston 2017): i.e., that a large animal charging at a group of smaller animals is usually followed by the scattering of the latter in flight. Furthermore, it's important to note that on our account the intentions are sensorimotor images all the way up and down the hierarchy of intention, and propositional attitudes only play a role in delimiting the transitions which may occur between these images. In Dretske's (2010) terminology, propositions are "structuring" causes and not "triggering" causes.

One may object nonetheless that this is a version of the causal theory after all, since propositional attitudes are exerting *some* causal influence over decisions. To this we believe the correct response is just to draw attentions to the details of our account. What we are interested in establishing is that intentions can be understood as mental images (not propositional attitudes), and that the mechanisms for bringing about action should not be understood on the model of the practical syllogism. If that is nonetheless to be labelled a causal theory of action, then in our view it is so distinct from the orthodox causal theory in its details as to make the shared label misleading. The CTA as ordinarily pitched is committed to much more than the bare thesis that propositional attitudes play *some* sort of causal role in action.

3.3 *Can Mental Imagery Be Abstract?*

Intentions can specify actions at a very high level of abstraction. Someone can have an intention to succeed in their career, to find true love, or to make the world a better place. Such a level of abstraction seems to require propositional content. That is because an intention which is so abstract is compatible with very many sorts of imagery. What would the imagery be that corresponds to the intention to make the world a better place? That is an abstract idea that does not seem to correspond to any specific image whatsoever. However, it is ideally represented in a propositional format, and we can easily see how an intention could propagate this abstract format into more concrete and intentions proximal to action if it were conceptual in form.

However, it is a mistake to assume that mental imagery cannot be abstract. Consider Block's (1983) discussion of the "photographic fallacy" in debates over mental imagery. One objection posed to theorists of mental imagery is based upon the idea that imagery-style representations must be determinate with respect to all their details. For example, Pylyshyn (1978) discusses the fact that when four-year olds are shown and then asked to draw a tilted beaker containing coloured water, they do not draw the water as remaining horizontal but rather draw it as perpendicular to the beaker (as if it did not respect gravity). He claims that this result motivates for the idea that visual experience is

"descriptive" (by which he means quasi-linguistic) rather than "pictorial," since if the visual representation in memory were pictorial then it should specify visual details determinately, in this case, the angle between the surface of the liquid and the beaker.

In response to this line of argument, Block replies that there is no need that pictorial representations must be determinate with respect to every visual feature that they represent. Sometimes, they can represent the existence of some determinable property (e.g., the way that mountainous terrain is represented on a map) while being non-committal on the specifics of the property (e.g., the number of mountains, and their specific topography). What sets the image apart from a proposition or conceptual content is rather that it is iconic in its format.[5] It lacks syntactic structure of the kind that language possesses. To make this difference stark, note that there is a syntactically correct way to decompose sentences. It does not do to split the sentence "Bob is happy" right down the middle, since it will make nonsense of the parts "Bob i" and "s happy." By contrast, an iconic representation (like a map) can be decomposed in any of number of ways without respect for any sort of syntax, and the parts will continue to represent just as well as the whole (Fodor 2007; cf. Quilty-Dunn 2016).

With this in mind, we are happy to put ourselves on the line in claiming that there can be intentions representing actions at a very high level of abstraction. We make sense of this in terms of a hierarchy of parameters, and for which specifications of values for the higher parameters will automatically activate a range of values at the lower levels in the hierarchy, but will leave them more or less unspecified within this range. For example, one may have a mental image of a drinking glass that is indeterminate with respect to the precise geometric dimensions of the glass. However, there will be a number of dimensions that will be primed on account of their likelihood relative to others. If prompted, we can make these dimensions more determinate, and presumably some similar process is at work for unconscious mental images, as well. We contend that the making determinate of lower level parameters corresponds to the means-end reasoning that is characteristic of practical reason once an agent is faced with implementing an abstract intention in a specific situational context.

It may be difficult to see how a picture would work like this in specific cases, like that of the intention to "make the world a better place." We think the right characterisation of these will need to proceed on a case by case basis. In this

5 Note that this sets our account apart from that of Mylopoulos and Pacherie (2017); although they also talk in terms of determinate and determinable properties, they take the representation of determinable properties by intentions to be propositional while we take them to be iconic or quasi-perceptual.

particular case, the content seems to be so vague that very many things come under the description. In fact, we would suggest that even on a propositional view of intention it will be difficult to explain what the content of this intention is. On our view, it is likely that there is no particular sensorimotor content that is appropriate to this content, but it also seems that if the agent were to deliberate on the meaning of this sentence then it would eventually result in a more concrete set of plans which could constitute a bona fide image-as-intention, on our view.

In any case, our main aim in this subsection is to dispute the idea that mental images cannot be abstract per se. That is a prejudice, since mental images can specify an indeterminate set of parameters or higher order variables that range over multiple parameters without ceasing to be quasi-perceptual in format. We can see the transition from highly abstract (or "distal") intentions to "proximal" intentions implementable in a concrete situation in terms of the progressive filling out of the lower-level parameters in a hierarchy, until such time as the image is of an appropriate grain to drive the decision variable to bound in the diffusion to bound mechanisms discussed in Section 2.

4 Conclusion

Diffusion-to-bound models of decision making are able to accurately model performance and response time in a wide range of perceptual decision-making paradigms. There is also very strong electrophysiological support of a model of this general kind. It is therefore a good place to begin in understanding the neural mechanisms which underlie decision-making in general.

Some theorists will demur from this assessment, on the grounds that intentional action has a character wholly different to the kinds of tasks in experimental paradigms modelled using diffusion-to-bound. In the dot motion paradigm, for example, the subject is required to make a decision about whether the dots are moving left or right. But in the cases that are paradigmatic in the philosophical literature, an agent is required to choose between two more courses of action on the basis of competing *reasons*. In some cases, for instance, the choice is between a course of action that is pragmatically best and one that is morally best (e.g., the choice to help someone at some cost to oneself). So one may refuse the relevance of diffusion-to-bound models on the basis that they seem to lack the rational component of choice that philosophers want to understand.

However, this ruling of diffusion-to-bound mechanisms as non-rational in principle seems to us premature. There are good reasons why the extrapolation

of diffusion-to-bound modelling ought to be attempted in spite of the superficial differences between intentional and perceptual decision-making. Primarily, to posit mechanisms distinct from those that are well evidenced in simpler contexts is to be committing oneself to a dual-process account on which different mechanisms are implemented for different sorts of decision making (Evans & Stanovich 2013). This is not an invalid path to take, but it comes with theoretical costs.

Primarily, positing multiple mechanisms obliges us to explain how those mechanisms interact in order to produce behaviour. With respect to intentional action, this explanation will need to be twofold. Firstly, it will need to explain how it is that one or another mechanism for action is selected over the other(s) that are posited by the account, and in what circumstances; secondly, it will need to explain how the representational format that is putatively appropriate to higher-level intention interfaces with the representational format that is putatively appropriate to bodily movement (Burnston 2017). Our account sidesteps these issues.

A second cost to this approach is that it risks indulging in exceptionalism about human action that is at odds with the view that we occupy a non-exceptional place in the natural world. Although it is true that human behaviour appears to have a remarkably wide scope and flexibility, it seems more plausible to think that this is the result of progressive iterations on simpler mechanisms that we share with other animals, rather than a discontinuous leap from those mechanisms. These costs, while clearly not fatal to the dual-mechanisms approach or to CTA, seem to us to justify the removal of objections against alternative approaches.

References

Anscombe, G.E.M. (1957). *Intention*. Oxford: Blackwell.

Block. N. (1983). The photographic fallacy in the debate about mental imagery. *Nous*, 17, 651–661.

Bratman, M. (1987). *Intentions, Plans, and Practical Reason*. Cambridge: Harvard University Press.

Buekens, F., Maesen, K., & Vanmechelen, X. (2001). Indexicaliteit en dynamische intenties. *Algemeen Nederlands Tijdschrift voor Wijsbegeerte*, 93, 165–180.

Burnston, D.C. (2017). Interface problems in the explanation of action. *Philosophical Explorations*, 20, 242–258.

Butterfill, S.A., & Sinigaglia, C. (2014). Intention and motor representation in purposive action. *Philosophy and Phenomenological Research*, 88, 119–145.

Cisek, P., & Kalaska, J.F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience*, 33, 269–298.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Davidson, D. (1963). Actions, reasons, and causes. Reprinted in *Essays on Actions and Events*, 1980. Oxford: Clarendon Press.

De Lafuente, V., & Romo, R. (2005). Neuronal correlates of subjective sensory experience. *Nature Neuroscience*, 8, 1698–1703.

Dretske, F. (2010). Triggering and structuring causes. In T. O'Connor & C. Sandis (eds.), *A Companion to the Philosophy of Action*, ch. 18. Oxford: Wiley-Blackwell.

Duque, J., Greenhouse, I., Labruna, L., & Ivry, R.B. (2017). Physiological markers of motor inhibition during human behaviour. *Trends in Neurosciences*, 40, 219–236.

Evans, J.S., & Stanovich, K.E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspectives on Psychological Science*, 8, 223–241.

Fodor, J.A. (2007). The revenge of the given. In B.P. McLaughlin & J.D. Cohen (eds.), *Contemporary Debates in Philosophy of Mind*, pp. 105–116. Malden: Blackwell.

Frith, C.D., Blakemore, S.-J., & Wolpert, D.M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London B*, 355, 1771–1788.

Gold, J.I., & Shadlen, M.N. (2002). Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36, 299–308.

Gold, J.I., & Shadlen, M.N. (2003). The influence of behavioural context on the representation of a perceptual decision in developing oculomotor commands. *Journal of Neuroscience*, 26, 632–651.

Gold, J.I., & Shadlen, M.N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30, 535–574.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Hommel, B. (2004). Event files: feature binding in and across action and perception. *Trends in Cognitive Sciences*, 8, 494–500.

Hume, D. (1739). *A Treatise of Human Nature*. Oxford: Clarendon Press, 1978.

Kahneman, D. (2002). Nobel Prize lecture: maps of bounded rationality, a perspective on intuitive judgment and choice. In T. Frangsmyr (ed.), *Nobel Prizes 2002: Nobel Prizes, Presentations, Biographies, & Lectures*, pp. 416–499. Stockholm: Almqvist & Wiksell Int.

Mele, A.R. (1992). *Springs of Action: Understanding Intentional Behaviour.* Oxford: Oxford University Press.

Millikan, R.G. (1995). Pushmi-pullyu representations. *Philosophical Perspectives*, 9, 185–200.

Murakami, M., Vincente, M.I., Costa, G.M., Mainen, Z.F. (2014). Neural antecedents of self-initiated actions in secondary motor cortex. Nat. Neurosci., 17, 1574–82.

Mylopoulos, M., & Pacherie, E. (2017). Intentions and motor representations: the interface challenge. *Review of Philosophy and Psychology*, 8, 317–336.

Nanay, B. (2016). The role of imagination in decision-making. *Mind & Language*, 31, 127–143.

Nanay, B. (2017). Multimodal mental imagery. *Cortex*, http://dx.doi.org/10.1016/j.cortex.2017.07.006.

O'Shaughnessy, B. (1980). *The Will: A Dual-Aspect Theory*, vol. 2. Cambridge: Cambridge University Press.

Pacherie, E. (2008). The phenomenology of action: a conceptual framework. *Cognition*, 107, 179–217.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, 9, 129–154.

Pylyshyn, Z. W. (1978). Imagery and artificial intelligence. from C.W. Savage, ed., Perception

Quilty-Dunn, J. (2016). Iconicity and the format of perception. *Journal of Consciousness Studies*, 23, 255–263.

Ratcliff, R., & Hacker, M.J. (1981). Speed and accuracy of same and different responses in perceptual matching. *Perception and Psychophysics*, 30, 303–307.

Roitman, J.D., & Shadlen, M.N. (2002). Response of neurons in the lateral interparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, 22, 9475–9489.

Schurger, A., Mylopoulos, M., & Rosenthal, D. (2016). Neural antecedents of spontaneous voluntary movement: a new perspective. *Trends in Cognitive Sciences*, 20, 77–79.

Schurger, A., Sitt, J.D., & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *PNAS*, 42, E2904–E2913.

Searle, J. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Shadlen, M.N., Hanks, T.D., Churchland, A.K., Kiani, R., & Yang, T. (2006). The speed and accuracy of a simple perceptual decision: a mathematical primer. In K. Doya et al. (eds.), *Bayesian Brain: Probabilistic Approaches to Neural Coding*, pp. 209–237. Cambridge: MIT Press.

Shallice, T., Burgess, P.W., Schon, F., & Baxter, D.M. (1989). The origins of utilisation behaviour. *Brain*, 112, 1587–1598.

Van Leeuwen, N. (2013). The meanings of "imagine" part I: constructive imagination. *Philosophy Compass*, 8, 220–230.

CHAPTER 2

Dual-System Theory and the Role of Consciousness in Intentional Action

Markus Schlosser

1 Introduction

The contemporary philosophy of action has revolved around the notion of intentional action, and it is widely agreed that intentionality distinguishes genuine action from mere behavior and mere happenings. In cognitive and social psychology, human cognition and agency are now widely explained in terms of the workings of two distinct systems (or types of mental processes). System 1, as it is often called, is characterized by processes that are fast, effortless, automatic, and unconscious. Examples include intuitive judgments, the recognition of faces and facial expressions, emotional reactions, fight-or-flight responses, and overlearned routines (such as typing a word, playing an instrument, driving a car, and so on). System 2 is engaged in processes that are slow, deliberate, controlled, and conscious. Paradigmatic examples are conscious and deliberate reasoning, the solution of novel and difficult problems, and the exercise of self-control. This dual-system framework has been deployed successfully in many areas of empirical research and I will assume, here, that it is on the right track in providing an empirically adequate account of human cognition and agency. (For an extensive overview and review see Evans 2008; see also Sloman 1996, Kahneman 2011, Evans & Stanovich 2013.)

Interestingly, the notion of intentionality does not occupy a theoretical role in the dual-system framework. Occasionally, however, some researchers in this area identify intentional actions with actions that are consciously controlled. Further, in the empirical literature it is often claimed that most of our behavior is driven by System 1 processes that are automatic and not consciously controlled. Given this, we face obvious and pressing questions. Is the philosophical account of intentional action compatible with the dual-system theory? If so, how can the dual-system theory account for intentional action? And what is the role of consciousness? If we simply identify intentional actions with consciously controlled System 2 outputs, we face the unpalatable conclusion that most of our behavior is not intentional, provided that most of our behavior is automatic and driven by System 1, as claimed in the empirical literature. We

- 978-90-04-40996-5

face, then, the further question of *how much* of our behavior can qualify as intentional within the dual-system framework.

In order to address those questions, we will need to get clearer about the role of consciousness in the dual-system theory and in intentional action. First, I will outline the standard view of intentional action, the dual-system theory, and the role of consciousness in the dual-system-theory. In order to discuss the role of consciousness in intentional action, I will consider five cases in which the activation of a social stereotype influences behavior in different ways and with varying degrees of conscious awareness. This will allow us to see how intentional action can be captured within the dual-system theory, and it will allow us to see that most of our everyday behavior can qualify as intentional, even if most of it is automatic. In particular, on the basis of the discussion of those five cases, I will propose a revised version of the standard view according to which automatic action (or so called "automatic goal pursuit") can qualify as *derivatively* intentional, if it has an appropriate history of habit formation.[1] An important general lesson will be that philosophical accounts of intentional action need to consider the role of consciousness in action more carefully, both in the guidance of action and in the history of habit formation. Further, the discussion will point to an important distinction between two kinds of intentions, and we will see how intentional action and goal-directed behavior can come apart.

2 The Standard View of Intentional Action

In the philosophy of mind and action, it is generally agreed that intentionality is the mark of genuine agency. For some time, it was also generally agreed that intentional agency can be explained in terms of the roles of the agent's desires and beliefs (largely due to the influence of Davidson 1963). In particular, it was widely assumed that to act intentionally is to act for a reason, and that acting for a reason is to be explained in terms of causation and rationalization by the agent's desires and beliefs. It is still widely held that there is a close connection between intentional action and acting for reasons—that intentional actions are *usually* performed for reasons. But the underlying claim that intentions can be reduced to desires and beliefs is now widely rejected (largely due to the influence of Bratman 1987). According to most contemporary versions of this standard theory of action, intentions play a crucial and irreducible role in practical reasoning, long-term planning, and in the initiation and guidance of

1 For related but nevertheless quite different approaches to this complex of issues see, for instance, Carruthers 2009, Frankish 2009, Evans 2010, and Hommel 2017.

action. On this view, the intentionality of action is to be explained in terms of the initiation and guidance by *intentions*, construed as irreducible mental states (Bratman 1987, Mele 1992, Enç 2003). In what follows, I will assume this as the current version of the standard view.

Philosophers of action have tended to assume that most of our everyday behavior qualifies as intentional, partly because they have assumed that the standard view can accommodate the intentionality of habitual actions. Davidson, for instance, noted that "we cannot suppose that whenever an agent acts intentionally he goes through a process of deliberation or reasoning" (1978: 85). On his view, actions are intentional only if they are caused and rationalized by the relevant desires and beliefs. But it is not required that the agent consciously considers the relevant desires and beliefs in reasoning. Rather, "if someone acts with an intention, he must have attitudes and beliefs from which, had he been aware of them and had the time, he *could* have reasoned that his action was desirable" (1978: 85). Davidson's example is an agent who adds spice to a stew with the intention of improving the taste. We can certainly imagine that this action is highly habitual, such that the agent's mind is preoccupied with something else. The action would nevertheless seem to be intentional (more on this below, in Section 6). Proponents of the view that intentions are irreducible usually offer a similar qualification. According to the current standard view, intentions are defined in terms of their functional roles, which include, most importantly, the initiation and guidance of action (Bratman 1987, Mele 1992, Enç 2003). According to the corresponding qualification, it is not required that the agent is consciously aware of the intention. It is sufficient, that is, if the action is initiated and guided by an intention that is consciously *accessible* (see Mele 2009: Ch. 2).

With this qualification in place, the standard view can accommodate the intentionality of habitual actions. For, even if most of our everyday behavior is habitual, as it seems, it is plausible to assume that most of our behavior is intentional, because it is plausible to assume that most habitual actions are initiated and guided by intentions that are consciously accessible (such that the agent could tell us his or her intention, if we asked).

3 Dual-System Theory

Dual-system theories "abound in cognitive and social psychology" (Evans 2008). They were first proposed to account for biases in logical reasoning, but are now widely deployed in the explanation of human cognition, decision-making, and agency. There are numerous versions of this view, but they share the same basic structure. Until recently, this structure was usually described by

means of dichotomies. System 1 processes were characterized as fast, effortless, automatic, and unconscious. System 2 processes were characterized as slow, deliberate, controlled, and conscious. More recently, this way of capturing the core of the dual-system theory has come under attack (for a summary of those criticisms and references see Evans & Stanovich 2013). Proponents of the view have acknowledged some of the criticisms, and they have adjusted the presentation of the view accordingly. In what follows, I will base my discussion on the recent accounts of the theory provided by Evans & Stanovich 2013 and Kahneman 2011.

According to Evans & Stanovich, the mentioned dichotomies capture "correlates" of the two systems, but they should not be taken as definitive (necessary and sufficient conditions). On their current view, the defining features of System 2 processes are "cognitive decoupling" and the "strong loading on the working memory resources that this requires" (2013: 226). This is in line with Kahneman's view, in which System 2 processes are defined in terms of the kind of effort that is required in the simultaneous maintenance of several (and therefore decoupled) representations in working memory, and that is characteristic of conscious problem solving, reasoning, and deliberation (2011: Ch. 2). Evans & Stanovich define System 1 now solely in terms of "autonomy". By this they mean, roughly, that System 1 responses are triggered by their stimulus without the involvement or intervention of System 2 (2013: 236). This coheres, again, with Kahneman's account, according to which the defining and unifying feature of System 1 processes is their automaticity (2011: Ch. 1).[2]

How do the two systems lead to action? The emerging consensus is that System 2 is endowed with the top-down control to inhibit or override System 1 processes. This is what Evans & Stanovich call a *default-interventionist* architecture.[3] On their account, System 1 usually provides the "default response" to a given task or situation, which arises quickly and intuitively. System 2 may intervene by overriding or inhibiting the default response, if there is sufficient time and if the agent is motivated to engage in effortful System 2 processing.

2 Evans & Stanovich (2013) now prefer the terminology of dual-*processes*, because they do not mean to imply that there are two separate systems in the brain. System 1, in particular, is not one unified mechanism, but an array of automatic processes or modules. Kahneman acknowledges this, but he nevertheless keeps the terminology of systems. This is common practice, and I will continue to use the systems terminology as well.

3 Evans & Stanovich distinguish between two architectures: *parallel-competitive* and *default-interventionist*. According to the former, the two systems simply compete for control, in parallel, and there is no systematic interaction. Evans & Stanovich argue convincingly that this is implausible (2013: 237). On this view, it seems mysterious how System 2 can ever gain control over behavior, because System 1 is always quicker than System 2. Further, System 2 operates with precious working memory resources, and one would expect that such a high-level system is provided with a more systematic access to the motor control system.

On their view, "most behavior will accord with defaults, and intervention will occur only when difficulty, novelty, and motivation combine to command the resources of working memory" (2013: 237). Kahneman proposes a similar architecture. On his view, System 1 continuously and automatically generates "suggestions" (impressions, desires, and intuitions), which are often endorsed by System 2 "with little or no modification" (2011: 24). System 2 engages in effortful processing and takes over when difficulties arise, when errors are detected, or when System 1 fails to provide a default response. In what follows, I will assume that a default-interventionist architecture provides the correct account of how the two systems lead to action, and the details of this view will be further specified and qualified in due course.

4 An Obvious but Unsuccessful Proposal

From what we have seen so far, it seems already clear that the standard view of intentional action and the dual-system theory are not incompatible. According to one obvious proposal on how to locate or capture intentional action within the dual-system framework, intentional actions are simply System 2 outputs: actions that are generated by System 2 processes. It seems that some of the researchers who work on automaticity and dual-system theory hold this view of intentional action. It seems that way, because automatic processes are sometimes *contrasted* with processes that are consciously controlled *and* intentional (see Bargh 1994, Bargh & Chartrand 1999, Bargh 2005, Evans & Stanovich 2013, for instance).[4] As mentioned, the reduction of intentions to desires and beliefs is now widely rejected in philosophy, but it is still commonly assumed that intentions are usually based on desires. One could capture this in accord with the view just outlined by adding the assumption that System 1 generates desires that may or may not be endorsed by System 2, assuming, further, that the endorsement of a desire may consist simply in the formation of an intention to pursue the desired goal.[5]

However, as already indicated, this view has a serious drawback. In the empirical literature, it is often claimed that the great majority of our behavior is automatic. In an influential review article, Bargh & Chartrand (1999) evoked

4 According to Bargh 1994, the absence of intentionality is one of the "four horsemen" of automaticity. In this relatively early review of automaticity research, Bargh characterized intentionality in terms of conscious control. In later work, Bargh mentions intentionality only in passing, usually as synonymous with conscious control.

5 The endorsement of a desire may of course involve more than that, such as the reflective judgment that the desired end is desirable or good. The proposed claim here is that endorsement *may* consist only in the formation of the corresponding intention.

the "unbearable automaticity of being", referring to Baumeister et al. (1998), who estimated that only about five percent of our behavior is consciously controlled. Similarly, Aarts et al. (2005: 466) confidently assert that "most of our social behavior occurs in an automatic fashion and originates in the unconscious". If we identified intentional actions with System 2 outputs, and if we contrasted thereby automatic System 1 outputs with intentional System 2 outputs, we would have to conclude that the great majority of our actions are not intentional. Moreover, the implicit identification of intentional actions with consciously controlled actions is at odds with the mentioned qualification of the standard view, according to which intentional action does not require conscious awareness of the relevant mental attitudes (see Section 2). And this means, in turn, that one could not accommodate the intentionality of habitual actions if one simply identified intentional actions with System 2 outputs.

Further, it is questionable that all desires are generated by System 1 and that the endorsement of desires is always due to System 2. Philosophers have long noted that many of our desires appear to be reason-responsive, and there is now a considerable amount of empirical evidence in support of this view (see Dill & Holton 2014). Some desires simply vanish when one learns that the desired object is unattainable, or when one judges that something else would be better, for instance. Some desires, that is, are responsive to the judgments and intentions generated by System 2, and some desires may even be generated by System 2. Given this, it would simply be a mistake to assume that *all* desires are first generated by System 1, and then, perhaps, endorsed by System 2.

Concerning the endorsement of desires, there is no good reason to think that this is always a System 2 process. System 2 processes are effortful in the sense that they involve cognitive decoupling and a strong load on working memory. It seems clear that we often endorse desires, intuitions, and other "default suggestions" consciously and swiftly, without engaging in effortful reasoning or cognition. Given this, the endorsement of desires need not be a System 2 process. Given, further, that the endorsement of a desire may consist simply in the formation of an intention to pursue the desired goal, this means that the formation of an intention need not be a System 2 process.

To see why this is plausible, note that many System 1 processes can be flexible, in the sense that their execution can be sensitive to the features of the particular situation and to the agent's beliefs about how to pursue or realize the goal. Examples from priming studies will help to illustrate this point. It has been shown, for instance, that when subjects are primed, in a first task, with words that are related to rudeness, kindness, helping, or so, they are more likely to engage in rude, polite, or helping behaviors in a subsequent task (Aarts

et al. 2005, Bargh & Williams 2006).[6] Rude, polite, or helping behaviors are typically sensitive to the features of the particular situation and to the agent's background beliefs about how to pursue the action in the given situation. This kind of flexibility cannot be explained in terms of rigid stimulus-response associations, which map a particular type of input (or situation) to a particular type of output (or action). For this reason, most researchers in this field agree that such actions are instances of "automatic goal pursuit": actions that are to be explained in terms of the activation of goal representations that initiate and guide the execution of the behavior (Custers & Aarts 2010).[7] On this view, the activation of the stimulus leads to the performance of the action indirectly, by way of the activation of the relevant goal, such that the resulting goal-pursuit is sensitive to the agent's beliefs about how that goal is to be pursued or realized in the circumstances. (Direct empirical evidence for this kind of flexibility is provided by Hassin et al. 2009.)

Now, according to one standard definition in psychology, goals are "internal representations of desired states" (Austin & Vancouver 1996: 338). According to another, a goal is the "representation of a future object that the organism is committed to approach or avoid" (Elliot & Fryer 2008: 244). The former corresponds, very roughly, to the philosophical notion of desire. The latter corresponds to the philosophical notion of intention, again very roughly.[8] We might say here that the definition of goals in psychology is ambiguous between the philosophical notions of desire and intention. Or, more positively, we might just as well say that the definition encompasses both the philosophical notion

6 Recently, some of the seminal experiments in this area have been called into question. Several attempts at replication have either failed or produced only significantly smaller effects. There is, however, also a very large body of research that corroborates the findings. Even if the effects are small and difficult to reproduce, it seems rather unlikely that there are no real effects that underlie the results in this area of research. For a recent discussion of those issues see Open Science Collaboration 2015.

7 According to Levy (2014: 118, note 5), the evidence does not indicate automatic goal pursuit, because the actions in question are merely *modulated* by the priming. This is a plausible interpretation of many of the classic findings. For instance, in one such experiment, priming influenced the accuracy of the task performance by influencing the speed of the task completion. But it can be argued that even in cases of this kind a non-conscious goal was "superimposed on an already activated parallel conscious task goal" (Bargh et al. 2001: 1024). In other experiments, an interpretation in terms of modulation is rather implausible. For instance, the subliminal priming of cooperation or helping behavior is not plausibly interpreted as a mere modulation (see Bargh et al. 2001, Aarts et al. 2005).

8 See Bratman 1987 and Mele 2009, for instance, who stress this element of commitment, or of having settled the question of which action to pursue, which distinguishes intentions from desires.

of desire and of intention. However, talk of automatic goal-*pursuit* suggests that the scientists in this particular research area have in mind the formation of a representational state that initiates and guides the pursuit of the relevant action—a state, in other words, that commits the agent to the pursuit of a particular action. Given this, we can say that the scientists in this area agree that automatic goal-pursuit, which is a paradigmatic System 1 process, is to be explained in terms of the automatic activation of *intentions*.

One might, of course, reject a definition of intentions solely in terms of the functional roles of initiation and guidance, and insist, for instance, that genuine intentions must be consciously formed or accessed. But this would not facilitate real progress. For if one simply claimed that intentional action requires conscious intention, without further qualification, one could not accommodate the apparent intentionality of habitual actions, and one would once again face the undesirable conclusion that most of our actions are not intentional.

We have seen that the standard view of intentional action is clearly not incompatible with the dual-system theory. But a satisfactory account of intentional action within this theory has to be more nuanced than the view considered in this section. In order to make constructive progress on this, we need to get clearer about the role of consciousness in the dual-system theory and in intentional action.

5 The Role of Consciousness in the Dual-System Theory

In the previous section, we have seen that consciousness is not a distinguishing feature of System 2, as System 1 processes may also involve consciousness. The default suggestions of System 1 appear often in consciousness, and if one endorses such suggestions by consciously forming a judgment or an intention without effortful reasoning, then the endorsement is itself part of the System 1 process. What is distinctive of System 2 processes is, more specifically, that one consciously conducts individual steps of reasoning in an effortful process that requires the simultaneous maintenance of decoupled representations in working memory. What kind of consciousness is at issue here?

In philosophy, it is common to distinguish between *phenomenal* and *access* consciousness. Phenomenal consciousness is usually characterized in terms of *what it is like* to have a certain experience (such as the distinctive quality of a certain visual or auditory perception). According to common characterizations of access consciousness, a representation is conscious if it is available for reasoning, decision-making, and the control of action (including verbal

report). Access consciousness is a functional notion—it specifies the functional roles of conscious mental states.

In a review article, Evans (2008) points out that there is an operational definition of consciousness that is shared, at least implicitly, by most dual-system theories: "System 2 thinking requires access to a central working memory system of limited capacity" and "what we are aware of at any given time is represented in this working memory, through which conscious thinking flows in a sequential manner" (2008: 259). This is a definition in terms of access and it specifies how representations become accessible for further processing—namely, by virtue of being in the central working memory system.

Evans notes, further, that this view is closely associated with the global workspace theory of consciousness. According to this view, a representation is conscious if it is "broadcast" in the "global workspace", which makes it available to a wide range of consumer systems (Baars 1997). Initially, this talk of broadcasting in the global workspace was merely a heuristic metaphor. But more recently this view has been developed into a global *neuronal* workspace theory of consciousness, which specifies concrete neural mechanisms for the workspace and for broadcasting. This theory is empirically well-motivated and it has been successfully deployed in the neuroscientific study of consciousness (Dehaene & Naccache 2001, Baars 2002). There has been some debate about whether or not the workspace can simply be identified with the central working memory system, but it is agreed that working memory is a central component of the workspace (Baars & Franklin 2003, Levy 2014). And as Evans notes, this "association of conscious thought with such a working memory explains the slow, sequential, and low-capacity nature of System 2" (2008: 259).

So, the role of consciousness in the dual-system theory is captured by an operational definition of consciousness that coheres with the philosophical notion of access consciousness and with the global (neuronal) workspace theory of consciousness. But can we plausibly restrict our considerations to access consciousness?

Note, first of all, that we are interested in the role of consciousness in the initiation and guidance of action. That is just to say that we are interested in the functional role of consciousness, which is precisely what a definition of access consciousness is supposed to capture. Note, moreover, that we are concerned here primarily with the roles of desires, beliefs, and intentions. To insist that phenomenal consciousness must play a role would commit one to the rather implausible view that conscious access to desires, beliefs, and intentions is always accompanied by phenomenal consciousness.

However, one may grant that access consciousness is all that matters in most cases and hold that there are other cases in which mental states influence

or motivate behavior in virtue of their phenomenal quality. Suppose, for instance, that certain aesthetic or moral judgments motivate actions in virtue of the fact that they are based on aesthetic perceptions or moral sentiments with certain phenomenal qualities. Would this not show that phenomenal consciousness can play an important and irreducible role in action?

What is at issue in such cases is always the "functional correlate" of the phenomenal experience—that is, the functional role that correlates with the phenomenal quality in question (see Chalmers 1995). It may well be that a mental state has a particular functional role in virtue of having a certain phenomenal quality. But its role in the initiation and guidance of action is still its functional role. Note that there are two different explanatory relations in play here. A mental state plays a role in the initiation and guidance of action *in virtue of* being a mental state with that functional role, and it may have and play this functional role *in virtue of* having a certain phenomenal quality. But its role in action remains its functional role. The explanatory relation between the phenomenal quality and the action is *indirect*, mediated by its functional role, which is the functional correlate of the phenomenal quality in question.

Note, finally, that the functional role of a conscious intention is by no means exhausted by the initiation and guidance of action (including verbal report). In particular, by virtue of being a *conscious* intention it has the functional role of being available for further deliberation and decision-making, and it plays this functional role by virtue of being broadcast in the global workspace (or central working memory system).

6 The Role of Consciousness in Intentional Action

Our main question is how to account for intentional action within the dual-system theory, and we have seen that we need to get clearer about the role of consciousness. As mentioned, in Section 2, it is common to qualify the standard view with the claim that intentional action does not *require* conscious deliberation or conscious awareness of the relevant intention. The fact that this is usually *added* as a qualification suggests that it is usually taken for granted that paradigmatic instances of intentional action *are* initiated and guided by conscious intentions (and, perhaps, reasons). I share this assumption, at least as a starting point for the present investigation.[9] In the previous section,

9 This assumption is in line with commonsense or folk intuitions about intentional action. The empirical evidence provided by Malle & Knobe 1997 and Malle et al. 2000 suggests that, according to the folk concept, an action is intentional and based on reasons only if the agent is aware of the relevant intention and reasons.

we have seen that the role of consciousness in the dual-system theory is captured by the global (neuronal) workspace theory of consciousness, and I have argued that the implicit restriction to access consciousness is unproblematic. With this framework in place, we can now address our main question in a more systematic and nuanced manner. We will consider a series of five cases, in which behavior is influenced by the activation of a social stereotype in five different ways, with varying degrees of consciousness. This distinction between five cases is not exhaustive. But it will cover the full range from automatic and unconscious goal pursuit to conscious and deliberate action, and it will thereby allow us to consider the full range from full System 1 to full System 2 engagement.

As mentioned, experiments show that behavior can be influenced by priming with words that are associated with concepts such as rudeness, politeness, or helping. Likewise, it has been shown that behavior can be influenced by priming with words that are associated with social stereotypes concerning race, gender, or social class, and the same has been shown for the activation of such stereotypes by the presence of group members and perception of group features, such as skin color (Bargh & Chartrand 1999, Aarts et al. 2005, Bargh & Williams 2006). Generally speaking, stereotyping involves generalization by way of categorization and association, and social stereotypes are commonly defined as "generalizations about the shared attributes of a group of people" (Judd & Park 1993). For instance, a person is perceived as Asian, female, or working class, due to certain superficial features, and this categorization is associated with features such as being good at math, bad at driving, or the like.[10] A stereotype is said to be *activated* when the perception or activation of one feature automatically activates an associated feature (or features). Experiments have shown that stereotypes can be activated without the subject's awareness, and they have shown that both the conscious and the unconscious activation of stereotypes tends to influence behavior. In the empirical literature, it is generally assumed that the influence on behavior may be mediated by one of the following two mechanisms (Bargh et al. 2001, Aarts et al. 2005, Bargh 2005, Bargh & Williams 2006). First, stereotypes may become associated with behavioral tendencies, such that the activation of a stereotypical feature automatically activates directly the associated behavioral tendency. Second, it is now assumed that stereotypes may also become associated with goal representations, such that the activation of the stereotype leads to behavior by activating the goal and subsequent goal pursuit. The main difference between the two mechanisms concerns behavioral flexibility. The first mechanism is

10 Social stereotypes need not be negative, and they need not be inaccurate. But they tend to be oversimplified overgeneralizations (Judd & Park 1993).

often described as rigid, in the sense that the activation of the stimulus directly activates a tendency to perform a certain type of behavior. The second mechanism is flexible, in the sense that it is sensitive to background beliefs about how to pursue or realize a goal in a given situation. To use the example from above, the activation of a social stereotype may become associated with the tendency to perform a certain type of rude behavior. Or, according to the second mechanism, the activation of the stereotype may automatically activate the goal to be rude, such that the pursuit of that goal is guided by background beliefs about how the goal is to be pursued in the given situation. The empirical evidence suggests that both mechanisms can operate entirely without conscious awareness. This evidence on stereotype activation is an important part of the evidence on implicit bias (Greenwald & Banaji 1995, Petty et al. 2008, for instance). The discussion of such cases will make it clear that an account of intentional action within the dual-system theory provides also a useful and plausible framework for how to interpret and diagnose the pernicious influence that implicit biases can have on our behavior.

Before we proceed, let me stress that the main purpose of considering the following five cases is to explore the various possibilities in the theoretical landscape, as it were—possibilities concerning the role of consciousness in intentional action and concerning the underlying mechanisms. In particular, the distinction between cases 4 and 5 will be based on the distinction between the two *possible* mechanisms of automatic goal pursuit just mentioned. I will provide references to empirical evidence, where possible, but it should be noted that all questions concerning the exact mechanisms and concerning whether, or to what extent, human agency instantiates any of those five cases are *empirical* questions—and most of them will remain *open* empirical questions for some time to come.

6.1 *Case 1: Full Awareness and Deliberate Action*

In this first case, the agent is aware of the stereotype activation and its influence leads to action by way of conscious deliberation. Suppose, for instance, that the agent encounters a member of a stereotyped group and that this automatically activates the associated stereotype. The activation of the stereotype is broadcast, in the global workspace (or working memory), and this instigates conscious deliberation about how to respond, in this situation. The deliberation is conducted by System 2 and the individual steps of the deliberative process are broadcast. This, we may assume, results in the conscious formation of an intention, which is then executed with conscious awareness.

In this case, the action is initiated and guided by a conscious intention, and this intention is based on conscious deliberation. Everyone would agree, I take

it, that the action in this case is clearly intentional in the fullest sense.[11] Note, though, that not all of the mentioned features are necessary to support the verdict that the action is intentional. According to the standard qualification of the standard view (see Section 2), initiation and guidance by an accessible intention is sufficient for intentional action. When we turn to case 3 we will see that this is in need of further qualification. But it is uncontroversial that initiation and guidance by a *conscious* intention is sufficient, and that conscious *deliberation* is not required. This is why it would be appropriate to describe the action in this case as intentional *and* deliberate or as intentional *in the fullest sense*.

Note that we do not assume that System 2 *intervenes* by inhibiting a default System 1 response. We may assume that System 2 takes over either because the agent is motivated to engage in deliberation, or because the stereotype activation fails to generate a default response. In either case, the core capacity of System 2 to conduct individual steps of reasoning in working memory is fully engaged, but the process is nevertheless not a *pure* System 2 process. In conscious deliberation, we consider reasons and we evaluate them in accord with certain rules (principles or normative standards). Typically, we are unaware of why or how we retrieve certain consideration as reasons and we are unaware of why or how we select the underlying rules. The relevant reasons simply appear in consciousness, and the relevant rules are usually operative in the background. The retrieval of reasons and the selection of rules has to be conducted by System 1, even in fully conscious deliberation. Otherwise, we would face a regress of consciously choosing reasons and rules, and consciously choosing reasons and rules for choosing *those* reasons and rules, and so on. Given this, it is clear that no cognitive process can be a pure System 2 process.

6.2 *Case 2: Full Awareness and No Deliberation*

Suppose now that the agent is aware of the stereotype activation, but that System 2 does not engage in conscious deliberation. As before, the stereotype is activated and its activation is broadcast (in the global workspace or working memory). But this time, the stereotype activation is not followed by conscious deliberation. Rather, it directly activates an associated goal representation (such as the goal to be rude, polite, or to engage in helping behavior). We may assume that the activation of the goal is also conscious, in the sense that it is

11 This verdict is supported by the standard view of intentional action, as outlined in Section 2, and by all its main rivals in the philosophy of action. Further, it is supported by the empirical evidence on the folk concept of intentional action provided by Malle & Knobe 1997 and Malle et al. 2000.

broadcast, and we may assume that the goal is endorsed by the conscious formation of an intention, which is then executed with conscious awareness.

The action is initiated and guided by a conscious intention. This intention is not based on conscious reasoning, although it may be based on reasons (which are accessible but not accessed in the situation). Either way, the action is clearly intentional, because it is initiated and guided by a conscious intention.[12]

This case does not involve genuine System 2 processing. The agent is aware of the stereotype activation and the goal activation, and the goal is endorsed by the conscious formation of an intention. But an engagement of System 2 would require, in addition, that individual steps of effortful reasoning are carried out in the central working memory system.

The action in this case is clearly intentional. I would suggest that our confidence in this verdict is to be explained, in part, by the assumption that the action is initiated and guided by a *conscious* intention. This will become clearer when we turn to the next case, where we will see that matters are far from straightforward once we remove the assumption that the action is initiated and guided by a conscious intention.

6.3 *Case 3: Goal Pursuit and Partial Automaticity*

As in case 2, the stereotype is activated and broadcast, and this automatically activates an associated goal. But suppose now that the goal activation is not broadcast. The agent, that is, is not aware of the goal activation and is therefore not in a position to endorse (or inhibit) the goal pursuit by forming a conscious intention. The empirical evidence suggests that such automatic goal activation may nevertheless result in automatic goal pursuit (for reviews see Bargh & Chartrand 1999, Bargh & Williams 2006, Custers & Aarts 2010). Let us suppose, then, that the goal activation initiates and guides an action without conscious awareness.

How should we judge this case? Is the action intentional? In this case, we need to know more. In particular, we need to know whether or not the performance of the action is habitual. To see this, return again to Davidson's example. The agent adds spice to the stew, mindlessly and without conscious awareness (as we are implicitly invited to assume). Why does it nevertheless seem that the action is intentional? The reason, I suggest, is twofold. First, we assume, with Davidson, that the relevant mental attitudes are accessible—if asked, the agent could readily declare his or her intention (to add spice in order to improve the taste). Second, we are clearly led to believe that the action is

12 As in case 1, this verdict is supported by the standard view, its main rivals in the philosophy of action, and the empirical evidence on the folk concept of intentional action provided by Malle & Knobe 1997 and Malle et al. 2000.

habitual and that the agent has the right history of habit formation. Clearly, the agent is not adding spice for the first time, and also not just for the second or third time. We would assume, rather, that the agent has done this over and over again. And we would assume that the agent did so on some occasions in the past with conscious intent.

With those background assumptions in place, we judge, readily, that the action is intentional. In particular, we implicitly assume that there is an underlying habit and an appropriate history of habit formation. There are, I suggest, two main types of such an *appropriate* history. Either the habit has been *formed* by several performances of the action with conscious intent. Or, it may be that the agent has acquired the habit in some other way, such as by imitation, but has later *endorsed* the action and its goal by forming a conscious intention. In either case, the intentionality of later manifestations of the habit is *derivative*: the intentionality of the action derives from earlier instances of acting with conscious intent.

Those considerations apply, *mutatis mutandis*, to our present case 3. If there is an appropriate history of habit formation, such that the habit was either formed or endorsed by acting on the stereotype activation with conscious intent, then a later manifestation, as in case 3, is derivatively intentional—provided, as we should add, that the intention is still consciously accessible.

The assessment of such cases becomes more difficult once we remove the assumption that there is an appropriate history of habit formation. Davidson's case would simply appear to be very odd without this assumption. Why on earth would one add spice to a stew automatically and without awareness if one had no corresponding history of habit formation? In case 3, in contrast, there is a possible explanation of how and why the agent acts automatically and without awareness. In the literature on implicit bias, it is often assumed that we may acquire the relevant stereotype-goal associations without a history of habit formation (or conscious endorsement), because it is assumed that we may acquire such associations by being exposed to them in our sociocultural environment. This would explain the association between the stereotype and the goal in our present case 3.

So, the stereotype is activated, and the agent is aware of this. But the agent is unaware of the goal activation and of the subsequent initiation and guidance of behavior. Is the action intentional? I am inclined to think that the action is not intentional.[13] But there is one complication that needs to be

13 As before, this intuition is in line with the standard view and with empirical evidence on the folk concept of intentions action (see footnote 11). Note that if an agent has acquired the stereotype-goal association by exposure, then the agent is probably unaware of the underlying association. That is, the agent is not only unaware of the goal activation in the

addressed here. We assume that the stereotype activation influences behavior by automatically activating a goal representation. A goal representation is, by definition, a mental state that can initiate and guide action. It has, that is, the functional role of an *intention* in the initiation and guidance of action. And this may be taken to suggest that the action is intentional. But what this shows, really, is that we need to be more careful, and that we need to make a distinction. We distinguished already between the automatic activation of a goal representation and the endorsement of such a goal by the conscious formation of an intention. This distinction is by no means *ad hoc*. One can find various incarnations of it in the empirical literature, and evidence from brain imaging studies suggests that the workings of automatic goal activations and conscious intentions are implemented by distinct regions in the brain (Frith et al. 2000, Bargh 2005, Pacherie & Haggard 2010, for instance). There is, that is, good reason to think that the distinction is real, but the terminology is optional. One may, for instance, reserve the term "intention" for mental states that have been consciously formed. Or one may introduce a technical term in order to distinguish between two *kinds* of intentions. For instance, in the empirical literature it is common to distinguish between conscious intentions and "motor intentions". That is nothing other than the distinction between conscious intentions and goal representations (sometimes also called "motor representations").

So, goal representations may be classified as forming a kind of intention. But, of course, this terminological decision should not lead one to conclude that all actions that are initiated and guided by goal representations are therefore intentional. What matters is not the terminology, but the two substantive issues that have already been discussed: Is the action habitual? Is the agent aware of the goal activation? Given this, we can uphold the proposed claim that the action in case 3 is not intentional, if it is not habitual, and if the agent is not aware of the goal activation and its influence on subsequent behavior.

Note, the suggestion is *not* that the performance of an intentional action requires initiation and guidance by a *conscious* intention. Rather, the suggestion is that if an action is initiated and guided by an automatically activated and unconscious goal representation, and if the action has no appropriate history of habit formation (or conscious endorsement), then the action is not intentional. In such cases, the action is goal-directed, but not intentional.

Note, finally, that System 2 is not involved in the *performance* of the action in any of the variations of case 3. But if the action has a history of habit formation (or conscious endorsement), then System 2 may have been involved

particular case. But the agent has probably never been aware of the fact that the activation of the stereotype activates an associated goal.

on past occasions, depending on whether or not those occasions involved effortful deliberation.

6.4 *Case 4: Goal Pursuit and Full Automaticity*

Assume now that the stereotype and the associated goal are activated automatically, and that neither the stereotype nor the goal is broadcast. As before, the agent is not in a position to endorse (or inhibit) the goal by forming a conscious intention. The empirical evidence suggests that the automatic activation of the goal may initiate and guide goal pursuit without conscious awareness (Bargh & Chartrand 1999, Aarts et al. 2005, Bargh & Williams 2006, Custers & Aarts 2010).

This is an example of fully automatic and unconscious goal pursuit, and the agent is unaware of the stereotype activation as well. As in case 3, we need to consider the agent's history in order to arrive at a judgment concerning intentionality. It seems, again, that if there is an appropriate history of habit formation or conscious endorsement, then the action is derivatively intentional, despite the fact that the agent executes the action in the present case without awareness. In support of this, consider the following. Suppose that when you drive to work in the morning, you take a right turn at the first junction. Sometimes when you approach the junction, you are aware of approaching a junction at which you have to turn right, and you then automatically prepare for taking the turn. On other occasions, you do not become aware of approaching a junction at which you have to turn right, but you automatically prepare for taking the turn all the same. In the first case, you are aware of the stimulus, in the second you are not. This is analogous to the difference between cases 3 and 4. In case 3 the agent is aware of the stereotype activation, in case 4 the agent is unaware. The analogy suggests that this difference does not make a difference concerning intentionality. In each of those four cases, the action is derivatively intentional if it is the manifestation of a habit with the right history (provided that the relevant intention is still accessible). And if there is no such history, the action does not seem to be intentional. Further, as in case 3, the action is goal-directed, but not intentional. System 2 is not involved in the performance of the action, but it may have been involved on past occasions, in the acquisition or endorsement of the relevant habit.

6.5 *Case 5: Full Automaticity and No Goal Activation*

As mentioned, the activation of a stereotype may influence behavior either by activating a goal or by activating more directly an associated behavioral tendency. Let us consider, then, a final case in which the automatic and unconscious activation of a stereotype influences behavior by way of automatically activating an associated behavioral tendency, and suppose that the agent is

unaware of all this (for empirical evidence see Bargh & Chartrand 1999, Aarts et al. 2005, Bargh & Williams 2006, Custers & Aarts 2010).

As in case 4, the agent is unaware of both the stereotype activation and its influence on behavior. And as in cases 3 and 4, we need to consider whether or not the action is the manifestation of a habit with the right history. And as the difference in the underlying mechanisms appears to be irrelevant, our assessment of intentionality should follow our discussion of cases 3 and 4. If the action is the manifestation of a habit with the right history, then it is derivatively intentional. If it is not habitual, then it is not intentional, because the agent is entirely unaware of being influenced by the activation of a stereotype. As in cases 3 and 4, System 2 is not involved in the performance of the action, but it may have been involved on past occasions, in the acquisition or endorsement of the relevant habit.

This case shows, though, that there is a sense in which an intentional action may not be goal-directed—namely, in the sense that it is not initiated and guided by a goal representation. But there is, nevertheless, derivative goal-directedness in such cases, because such actions must be derivatively intentional, if they are intentional at all.

7 The Standard View Revised

The discussion of those cases made it clear, I think, that the standard view of intentional action is in need of modification and further qualification. Consider, then, the following revised version of the standard view, which summarizes the conclusions and suggestions from the previous section (in a rough and ready fashion).

An action is intentional if and only if:

> Either (1) the performance of the action is initiated and guided by a conscious intention.
>
> Or (2) the action is derivatively intentional.
>
> An action is derivatively intentional if and only if:
>
> Either (2a) the action is the manifestation of a habit that has been formed by performing actions of this type with conscious intent (and this intention is still consciously accessible and would still be endorsed).
>
> Or (2b) the action is the manifestation of a habit that has been acquired in some other way and has later been endorsed by the conscious formation of an intention (and this intention is still consciously accessible and would still be endorsed).

According to the standard qualification of the standard view (see Section 2), intentional action does not require that the relevant mental attitudes are

consciously accessed—accessibility is sufficient. According to the proposed revision, conscious access is also not necessary, but accessibility is not sufficient. The agent must either have a conscious intention that initiates and guides the action (to satisfy 1), or the agent must have consciously formed the relevant intention at some point in the past (to satisfy 2a or 2b). So, according to this account, intentional action does depend on conscious intention, but the performance of particular intentional actions does not. The account makes no explicit mention of Systems 1 and 2. But the discussion of the five cases in the previous section made it clear why and how this account of intentional action can be captured within the dual-system theory (in conjunction with the global workspace theory of consciousness).

It is important to note that this account does not require that the agent has at any point the conscious intention to acquire the habit. The condition on habit formation (2a) requires that the habit is formed by performing the *action* with conscious intent. This requires, typically, that the action is performed or practiced repeatedly, but not that the agent has the conscious intention *to acquire the habit*. Note, further, that the account does not entail that the endorsement of a habit (in 2b) must itself be a mental action. The condition requires that the habit is endorsed by the formation of a conscious intention, and the formation of a conscious intention need not be an action. Arguably, the formation of an intention is an action only if it is motivated by a further desire or intention (such as the desire or intention to settle the practical question at hand). It seems perfectly possible that the formation of some intentions is not motivated in this way, and it seems perfectly possible that we may consciously acquire intentions in this passive or "non-actional" manner (for more on this see Mele 2003: Ch. 9).

8 Lessons and Conclusions

There is no straightforward way to locate intentional action in the dual-system theory, such as by identifying intentional actions with System 2 outputs. But we have seen that the standard view of intentional action can be captured within the dual-system theory, and it has emerged that doing so suggests plausible modifications and qualifications of the view. Further, we have seen that it is important to distinguish between goal representations and conscious intentions, or, alternatively, between two kinds of intentions.

An important lesson is that philosophical accounts of intentional action need to pay more attention to the role of consciousness in action. I have suggested that intentional action depends on consciousness. Consciousness, that is, must play a role at some point, either in the initiation and guidance of the

action or during the formation or endorsement of the relevant habit. This does not mean that intentional action depends on the involvement of System 2, because the conscious endorsement of a goal need not involve System 2 processing, as I have argued.

What can we say about *how much* of our behavior can qualify as intentional? We can conclude that most of our everyday behavior may well be intentional, even if most of it is automatic, because most automatic actions may well be habitual and derivatively intentional. This, I should note, includes another way in which automatic actions can qualify as derivatively intentional. Many automatic actions are sub-routines that are in the service of consciously accessible goals and intentions. Common examples are the sub-actions that one performs while playing an instrument or while driving a car. This kind of derivative intentionality is included, because the initiation and guidance of such sub-routines is always habitual.

To conclude, then, we have seen that the philosophical standard view of intentional action can be captured within the dual-system theory, and we have seen that doing so offers important lessons on how to think about the role of consciousness in intentional action. And we have seen that the findings from the empirical research on automaticity are not so "unbearable" after all, because they do not undermine the assumption that most of our everyday behavior can qualify as intentional.

Acknowledgements

A draft of this paper was presented at workshops and conferences at University College Dublin, University College Cork, UC Leuven, and the HU Berlin. Many thanks for the questions and comments that have helped to improve the manuscript, and special thanks to Francesca Bunkenborg for the very helpful commentary presented at the HU Berlin.

References

Aarts, H., Chartrand, T.L., Custers, R., Danner, U., Dik, G., Jefferis, V.E., & Cheng, C.M. (2005). Social stereotypes and automatic goal pursuit. *Social Cognition* 23: 465–490.

Austin, J.J., & Vancouver, J.B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, 120: 338–375.

Baars, B.J. (1997). *In the Theatre of Consciousness*. Oxford: Oxford University Press.

Baars, B.J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences* 6: 47–52.

Baars, B.J. & Franklin, S. (2003). How conscious experience and working memory interact. *Trends in Cognitive Sciences* 7: 166–172.

Bargh, J.A. (1994). The four horsemen of automaticity: Awareness, efficiency, intention, and control in social cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.). Handbook of social cognition (2nd ed.). Hillsdale: Erlbaum, pp. 1–40.

Bargh, J.A. (2005). Bypassing the will: Toward demystifying the nonconscious control of social behavior. In R.R. Hassin, J.S. Uleman & J.A. Bargh (Eds.). *The New Unconscious*. Oxford University Press, pp. 37–58.

Bargh, J.A., & Chartrand, T.J. (1999). The unbearable automaticity of being. *American Psychologist* 54: 462–479.

Bargh, J.A., Gollwitzer, P.M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology* 81: 1014–1027.

Bargh, J.A., & Williams, E.L. (2006). The automaticity of social life. *Current Directions in Psychological Science* 15: 1–4.

Baumeister, R.F., Bratslavsky, E., Muraven, M., & Tice, D.M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology* 74: 1252–1265.

Bratman, M.E. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Carruthers, P. (2009). An architecture for dual reasoning. In J.St.B.T. Evans & K. Frankish (Eds.). *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press, pp. 109–127.

Chalmers, D.J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2: 200–19.

Custers, R., & Aarts, H. (2010). The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science* 329: 47–50.

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, 60: 685–700; reprinted in D. Davidson (1980). *Essays on Actions and Events*. Oxford: Clarendon Press, pp. 3–20.

Davidson, D. (1978). Intending. In Y. Yovel (Ed.), *Philosophy of History and Action*, Dordrecht: D. Reidel; reprinted in D. Davidson (1980). *Essays on Actions and Events*. Oxford: Clarendon Press, pp. 83–102.

Dehaene S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* 79: 1–37.

Dill, B., & Holton, R. (2014). The addict in us all. *Frontiers in Psychiatry* 5: 1–20.

Elliot, A.J., & Fryer, J.W. (2008). The goal concept in psychology. In J. Shah & W. Gardner (Eds.), *Handbook of Motivational Science*. New York: Guilford Press, pp. 235–250.

Enç, B. (2003). *How We Act: Causes, Reasons, and Intentions*. Oxford: Oxford University Press.

Evans, J. St. B.T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* 59: 255–278.

Evans, J. St. B.T. (2010). *Thinking Twice: Two Minds in One Brain*. New York: Oxford University Press.

Evans, J. St. B.T., & Stanovich, K.E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science* 8: 223–241.

Frankish, K. (2009). Systems and levels: Dual-system theories and the personal-subpersonal distinction. In J.St.B.T. Evans & K. Frankish (Eds.). *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press, pp. 89–107.

Frith, C.D., S. Blakemore, & Wolpert, D.M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London B* 355: 1771–1788.

Greenwald, A.G., & Banaji, R.M. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* 102: 4–27.

Hassin, R.R., Bargh, J.A., & Zimerman, S. (2009). Automatic and flexible: The case of non-conscious goal pursuit. *Social Cognition* 27: 20–36.

Hommel, B. (2017). Consciousness and action control. In T. Egner (Ed.). *The Wiley Handbook of Cognitive Control*. Oxford: Wiley-Blackwell, pp. 111–123.

Judd, C.M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review* 100: 109–128.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Levy, N. (2014). *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

Malle, B.F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology* 33: 101–121.

Malle, B.F., Knobe, J., O'Laughlin, M.J., Pearce, GE., & Nelson, S.E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions. *Journal of Personality and Social Psychology* 79: 309–326.

Mele, A.R. (1992). *Springs of Action: Understanding Intentional Behavior*. Oxford: Oxford University Press.

Mele, A.R. (2003). *Motivation and Agency*. Oxford University Press.

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science* 349: acc4716–1–aac4716-8.

Pacherie, E. & Haggard, P. (2010). What are intentions? In L. Nadel & W. Sinnott-Armstrong (eds.), *Conscious Will and Responsibility: A Tribute to Benjamin Libet*. Oxford University Press, pp. 70–84.

Petty, R.E., Fazio, R.H., & Brinol, P. (Eds.). (2008). *Attitudes: Insights From the New Implicit Measures.* New York: Psychology Press.

Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin* 119: 3–22.

CHAPTER 3

When Do Robots have Free Will? Exploring the Relationships between (Attributions of) Consciousness and Free Will

Eddy Nahmias, Corey Hill Allen and Bradley Loveall

1 The Often Implicit, Yet Essential, Connection between Consciousness and Free Will

Imagine that, in the future, humans develop the technology to construct humanoid robots with very sophisticated computers instead of brains and with bodies made out of metal, plastic, and synthetic materials. The robots look, talk, and act just like humans and are able to integrate into human society and to interact with humans across any situation. They work in our offices and our restaurants, teach in our schools, and discuss the important matters of the day in our bars and coffeehouses. How do you suppose you'd respond to one of these robots if you were to discover them attempting to steal your wallet or insulting your friend? Would you regard them as free and morally responsible agents, genuinely deserving of blame and punishment?

If you're like most people, you are more likely to regard these robots as having free will and being morally responsible if you believe that they are conscious rather than non-conscious. That is, if you think that the robots *actually experience* sensations and emotions, you are more likely to regard them as having free will and being morally responsible than if you think they simply behave like humans based on their internal programming but with no conscious experiences at all. But why do many people have this intuition? Philosophers and scientists typically assume that there is a deep connection between consciousness and free will, but few have developed theories to explain this connection. To the extent that they have, it's typically via some cognitive capacity thought to be important for free will, such as reasoning or deliberation, that consciousness is supposed to enable or bolster, at least in humans. But this sort of connection between consciousness and free will is relatively weak. First, it's contingent; given our particular cognitive architecture, it holds, but if robots or aliens could carry out the relevant cognitive capacities *without* being conscious, this would suggest that consciousness is not constitutive of, or essential for, free will. Second, this connection is derivative, since the main connection goes through some capacity other than consciousness. Finally, this connection

- 978-90-04-40996-5

does not seem to be focused on *phenomenal* consciousness (first-person experience or qualia), but instead, on *access* consciousness or self-awareness (more on these distinctions below).

Perhaps the most substantive claims about the necessity of consciousness for free will come from scientists who discuss free will. For instance, "willusionists," who say that scientific research suggests that free will is an illusion typically reach that conclusion by assuming that free will requires that one's "conscious will" causes one's actions. Since willusionists argue that scientific research shows that conscious will does *not* cause our actions, they conclude that free will is an illusion (see, e.g., Libet 1999; Wegner 2002; Bargh 2008; and Harris 2012). Willusionists support their conclusion by arguing that research in neuroscience and psychology suggests that conscious mental states and processes do not play a causal role in decisions and actions, because non-conscious neural or psychological processes happen first.[1] These scientists, however, do not say much about *why* consciousness is crucial for free will. They typically assert the essential connection with claims such as Roy Baumeister's: "If there are any genuine phenomena associated with the concept of free will, they most likely involve conscious choice" (2008, 76; see also Libet 1999 and Wegner 2002).

Philosophers tend to agree that consciousness is necessary for free will. For instance, when they respond to willusionists, they typically dispute the relevance of the neuroscientific studies, the dualist definition of free will, or *which* conscious mental states are relevant to free choices (e.g., important deliberations, not consciousness of an intention to move a moment before an inconsequential movement). But these philosophers do not reject the importance of consciousness for free will. For example, Al Mele notes that "[i]f all behavior were produced *only* by nonconscious processes, and if conscious decisions (or choices) and intentions (along with their physical correlates) were to play no role at all in producing any corresponding actions, free will would be in dire straits" (2010, 43). And Eddy Nahmias suggests that: "Free will requires that one's actions properly derive from reasons for action that one has at some point consciously considered (or at least that one would accept if one considered them)" (2014, 18).

But it did not take scientists challenging the role of consciousness for philosophers to suggest that it is required for free will. For instance, Galen Strawson

1 See Mele (2010) and Nahmias (2014) for various responses to the evidence willusionists use, their interpretation of its relevance to conscious intention formation and to free will, and to the definition of free will the willusionists (mistakenly) assert as dominant in philosophy and as commonsensical.

writes, "To be responsible… one must have consciously and explicitly chosen to be the way one is, mentally speaking, in certain respects, and one must have brought it about that one is that way" (1994, 6). Randy Clarke writes, "Free will requires a capacity for rational self-determination… a free agent must be able to exercise [this capacity] consciously … an agent who is not even capable of conscious, effective practical reasoning does not have the variety of rational control in acting that we prize" (2003, 16). And Isaiah Berlin writes, "I wish to be a subject, not an object; to be moved by reasons, by conscious purposes, which are my own, not by causes that affect me, as it were, from outside" (1958, 203). Across diverging theorists—from compatibilists to libertarians to skeptics about free will—one truth seems to be self-evident: that free will requires consciousness.

Yet, despite the fact that this free will-consciousness connection is so pervasive among scientists and philosophers, the connection has typically been asserted without much explanation or defense, often taken for granted or left implicit. As Gregg Caruso points out, this is an explanatory gap that must be filled: "Clarifying the relationship between consciousness and free will is imperative if we want to evaluate the various arguments for and against free will" (2016, 78). It may be that the connection is under-analyzed precisely because it is so intuitive that we tend not to notice it. We've never encountered agents that seem autonomous and responsible that we did not also take to be conscious. But perhaps that will change if we develop autonomous robots (or meet intelligent aliens) and we are unsure about their consciousness. Furthermore, perhaps we can learn more about the free will-consciousness connection by exploring ordinary people's reactions to such possibilities and trying to tease apart which features of consciousness underlie their attributions of free will.

Indeed, until recently, little attention has been paid to ordinary people's attitudes about the connection between consciousness, free will, and moral responsibility. Given that many free will theorists appeal to commonly held intuitions as evidence for their theory, it is important that philosophical theorizing about concepts such as free will track ordinary understanding of those conceptions, or conversely, provide an error theory to explain why those intuitions are mistaken (see, e.g., Murray & Nahmias 2014). While some experimental philosophers and psychologists have conducted studies on people's intuitions and attitudes about free will and moral responsibility, the relationship between free will and consciousness has been largely underexplored.[2]

2 See, e.g., Nahmias, Morris, Nadelhoffer, & Turner (2006); Nichols & Knobe (2007); Monroe & Malle (2010); Monroe, Dillon, & Malle (2014); Stillman, Baumeister, & Mele (2011); Vonasch, Baumeister, & Mele (2018).

Recognizing this gap in the literature, Joshua Shepherd conducted a series of studies designed to understand people's attitudes about the role that consciousness plays in grounding free will and moral responsibility (e.g., 2012, 2015). Across several studies, Shepherd finds that people are much more likely to judge an agent to be free and responsible if the agent is conscious and to judge that an agent's particular actions are free and responsible when they are carried out consciously rather than non-consciously.

In one intriguing study, Shepherd asked participants to imagine the existence of sophisticated humanoids who "look, talk, and act just like humans, and they integrate into human society with no problem at all" (2015, 939). Some participants read scenarios that describe these creations as possessing consciousness: "They actually *feel* pain, *experience* emotions, *see* colors, and *consciously* deliberate about what to do"; while other participants read that the robots are *not* conscious: "they do not *actually feel* pain … they do not *experience* emotions, they do not *see* colors, and they do not *consciously* deliberate about what to do" (939). Some participants read a scenario in which one of these robots, Sal, steals a wallet he finds, while other participants read a scenario in which Sal returns a wallet he finds. Across scenarios describing both the bad and good action, participants who were told that the robots were conscious tended to judge Sal to be free, blameworthy (or praiseworthy), and morally responsible, while those who were told that the robots were not conscious tended *not* to attribute these features to Sal (940). Shepherd concludes that these results show that most people believe that conscious states and processes play a central role in grounding free will and moral responsibility. Shepherd speculates about some reasons people may make the connection between free will and consciousness, and how some philosophical theories might align with or revise ordinary intuitions about the connection. He concludes that his findings suggest that philosophers

> either develop a substantive theory of the connection between consciousness on the one hand and free will and moral responsibility on the other, or offer justification for jettisoning this seemingly central part of our commonsense understanding of free will and moral responsibility. (944)[3]

Here, we describe studies we conducted that build off of Shepherd's studies, and we take up his explanatory challenge. In fact, we hope to use our studies to begin to distinguish *which* features of consciousness, or the capacities they

3 See, e.g., Huebner (2010); Jack & Robbins (2012); Knobe & Prinz (2008).

might allow, people see as most essential for free will, and *why* these features or capacities are especially relevant or essential. Our aim is to bring to the surface implicit connections that might underlie the strong intuition among most people—including most philosophers and scientists who discuss free will—that the capacity to have conscious experiences is crucial for free will and responsible agency. If so, it might be that philosophers can even develop plausible theories by drawing on the connections underlying ordinary thinking. In any case, we'll try to develop one such theory that we take to be plausible.

2 Some Potential Connections between Consciousness and Free Will

There are several plausible features of consciousness that could be brought to bear on free will and moral responsibility. One historically prominent route emphasizes the phenomenology of free will. For example, Jean-Paul Sartre (1943) suggested that being conscious (or perhaps self-conscious) necessarily makes one radically free. Others have argued that the experience of free will is necessary for having free will. Galen Strawson, for instance, writes

> But why should lack of explicit awareness of [freedom] be supposed to have as a consequence lack of [freedom] itself, as lack of any sense or conception of U-freedom seems to have as a consequence lack of U-freedom itself? Well, that is the question. But it does seem to be so. (1986, 307)

In both cases the idea seems to be that the first-person experience of having open alternatives for future choices is essential for possessing free will. Yet, it's not clear *why* the experience of freedom is necessary for possessing free will. Indeed, this suggestion raises more questions than it answers. For example, must the phenomenal experience of freedom play some *causal* role in one's decisions and actions, or could the experience be epiphenomenal? If a causal role is required, then the question is what the experience of freedom is causing and what agents lack if they can behave in similar ways without the experience of freedom. If the experience plays no causal role, then it is even more mysterious what role it plays in making the agent free or morally responsible.

Another feature of consciousness that might be relevant to free will is its role in grounding libertarian free will. One might defend this view in a few ways. Perhaps, for example, consciousness bears some relation to a nonphysical mind or soul that can make free choices and causally interact with the physical brain and body (e.g., Swinburne 2013). The idea seems to be that

the conscious self can be an uncaused cause, free from the deterministic chain of cause and effect in the physical world. In addition to the mysterious causal interaction between non-physical minds and physical bodies that this view suggests, it also does not explain *why* it is the mind or soul's capacity for *consciousness* that allows it to be an uncaused cause. Other libertarians have connected consciousness to free will via quantum theory, gesturing towards the indeterminism of the dominant theory of quantum mechanics or towards the alleged role that consciousness plays in collapsing the wave function (see, e.g., Penrose 1991; Stapp 2001; and Hodgson 2002). At this stage, however, these views seem to try to solve the mystery of (libertarian) free will by conjoining the mystery of quantum physics with the mystery of consciousness. Robert Kane, offering a naturalistic libertarian view, suggests that consciousness may allow the unity of the self: "it may be that both the unity of conscious experience and the unity of the self-network are somehow related to the quantum character of reality" (1998, 195). It is plausible that free will requires a unified self and that we have conscious experiences of being a unified self at a time and across time (an experience some Humeans and Buddhists would say is illusory), and below we suggest there are specific features of conscious experience relevant to demarcating the self. It is unclear, however, why the unity of self should be associated specifically with a *libertarian* theory of free will.

Indeed, some compatibilists about free will and determinism suggest that consciousness is relevant because it allows the integration of information such that the agent has the ability to access her competing reasons and values during deliberation. For example, Neil Levy argues for the "consciousness thesis" which says that "consciousness of some of the facts that give our actions their moral significance is a necessary condition for moral responsibility (2014, 1).[4] However, like some other compatibilist theories described below, Levy's does not focus on *phenomenal* consciousness (qualia or the sensory and emotional qualities associated with conscious experiences). Rather, Levy focuses specifically on *access* consciousness. Information is "access conscious" to an agent when it is accessible for use by a wide range of cognitive systems, including deliberation, reasoning, and verbal report (Block 1995). It's controversial whether the distinction between these concepts of consciousness maps smoothly onto human psychology or fits with ordinary people's understanding

4 Note that Levy focuses here on moral responsibility, not free will, and that he is a somewhat non-standard compatibilist, in that he thinks determinism does not rule out free will or moral responsibility, but he thinks we have neither because of an argument from luck.

of consciousness, but Levy uses it to focus on the importance of access consciousness, suggesting that phenomenal consciousness is irrelevant.

Another prominent compatibilist theory may similarly suggest that, for the freedom associated with moral responsibility, it is accessibility of information to reasoning processes that matters more than phenomenal or qualitative experience. Reasons-responsive theories emphasize the control that access consciousness enables over one's decisions and actions as a result of reasoning, deliberation, and self-reflection. On Fischer and Ravizza's (1998) account, for example, agents are morally responsible for actions that are caused by moderately reasons-responsive mechanisms. As Shepherd (2015, 943) points out, however, reasons-responsive theories often emphasize features of decision-making that are unrelated to conscious experience. Indeed, on some interpretations of reasons-responsive compatibilism, it's not clear that agents must even be conscious in order to have the capacity for free will and moral responsibility (e.g., Yaffe 2012, 182). While access-conscious mental states are certainly required on Fischer and Ravizza's view, it's not entirely clear what role, if any, *phenomenal* mental states and processes play in enabling free will and moral responsibility. Though Fischer and Ravizza argue that free agents must "take ownership" of the relevant mechanisms, they don't say whether these agents must be conscious of the practical reasoning processes that they carry out.

Another type of compatibilist account suggests that free will involves decisions and actions caused by the "deep self" or "real self" as labeled by Susan Wolf (1990). She is referring to theories that pick out freely willed actions as the ones caused by those (first-order) desires that the agent (second-order) desires to move her (Frankfurt 1971), or that she identifies herself with (Frankfurt 1987), or that accord with her considered values (Watson 1975). These theories seem to require that the agent has free will only if she acts on motivational states that she is consciously aware of and consciously endorses. If so, they also seem to link free will to access consciousness or a type of self-reflective awareness. One might wonder whether such higher-order representational states require phenomenal consciousness—whether there must be anything it is like to experience them—or whether a sophisticated robot (or even humans in some instances) could carry out such higher-order representation without any phenomenology at all. As Daniel Dennett suggests in describing such robots: "our imagined [non-conscious] creatures would be equally able to engage in rational self-evaluation. They would be equipped to react appropriately when we represent reasons to them. Isn't that what freedom hinges upon, whether or not it amounts to consciousness?" (1984, 43).

Yet another type of compatibilist theory, related to these deep self views, suggests that consciousness is not required for free and responsible actions.

These "quality of will" (or self-expression) theories say that agents are responsible for those actions that express the agent's concern or consideration of others (their quality of will), which is sometimes identified as actions expressing the agent's deep self (see, e.g., Arpaly 2003; Smith 2005; Sher 2009; Buss 2012; and Sripada 2016). These theorists have argued that actions can express an agent's quality of will even when motivated by values that the agent consciously rejects (e.g., Huck Finn protecting Jim even though he thinks he should not) or in cases of negligence, when the agent, for instance, does not care enough to show up to help his friend move but did not *consciously* try to forget. These instances of responsibility for specific non-conscious actions are plausible. However, it's unclear whether these theorists would argue that it's irrelevant to free will and responsibility whether a creature is phenomenally conscious at all (sometimes called "creature consciousness").

Indeed, most quality of will theorists take their inspiration from Peter Strawson, who lays the foundation for the proposal that we will suggest for the connection between phenomenal consciousness and free will. Strawson (1962) argues that freedom and responsibility are grounded in our reactive attitudes, such as indignation, gratitude, and guilt, that we express in interpersonal relationships. According to Strawson, agents are morally responsible when they are apt targets of these reactive attitudes—for instance, when it is appropriate to feel resentment towards them when their actions express a poor quality of will towards you, to feel gratitude towards them when their actions express a good quality of will towards you, and to feel guilt when your own action expresses poor quality of will towards others. As such, Strawson's account ties free and responsible agency to the capacity to experience and express certain moral emotions, and it suggests that we attribute such agency only to other agents whom we perceive as feeling relevant emotions and expressing them in their actions. Hence, on a Strawsonian account, it might be the ability to consciously experience emotions that bridges phenomenal consciousness and free will.

A related view suggests that free and responsible agency is tied to our ability to *care* about what motivates us. On this view, actions expressing our deep self or quality of will are those that are caused by what we care about. For instance, Harry Frankfurt modifies his earlier views that focused on higher-order desire and identification to focus on the role that caring plays for grounding agency. He writes that a free agent is "prepared to endorse or repudiate the motives from which he acts … to guide his conduct in accordance with what he really cares about"; and he adds that what is essential to freedom pertains to "what a person cares about, what he considers important to him… Caring is essentially volitional… similar to the force of love" (1999, 113–114). Chandra Sripada

develops these ideas, arguing that one is morally responsible for an action only when it expresses one's deep self, and that the actions that express one's self are precisely those motivated by one's cares (2016). He defines cares in terms of what functional role they play in our psychology: they are foundational motives—i.e., intrinsic and ultimate, such that many of our other desires motivate actions that aim at satisfying our cares—and we desire to maintain our cares, and feel a sense of loss when we alter them.

Sripada's conative account is contrasted with the more cognitive deep self approaches described above that seem to require access (or self-) consciousness. His account suggests a more important role for *phenomenal* consciousness, because it is directly tied to emotion. He writes, "caring is also associated with a rich and distinctive profile of emotional responses that are finely tuned to the fortunes of the thing that is the object of the care" (2016, 1210).[5] For instance, if Paul *cares* about the plight of Sudanese children, then "if the fortunes of the Sudanese children are set back, Paul is susceptible to sadness, disapprobation, and despair" (1230–31). Now, Sripada and other theorists writing about cares do not explicitly point out that phenomenal consciousness is crucial for an agent to be able to have cares, likely because they take it to be understood that feeling emotions like sadness, despair, and joy requires the ability to have phenomenally conscious experiences. Indeed, it is difficult for us to imagine creatures (such as humanoid robots) that lack conscious experiences entirely yet are also able to have the sort of emotional responses required for them to *care* about what they do or what happens to them. They might have motivational states, they might represent them and evaluate them, but they would not seem to have the capacities to feel the sort of satisfaction or suffering that seem constitutive of caring.

We have now seen two related accounts that situate certain emotions at the heart of free and responsible agency, Strawsonian accounts based on reactive attitudes and self-expressive accounts that focus on the capacity to care.[6] We propose that these connections provide the link between free will and (specifically) phenomenal consciousness. Then, we offer some initial evidence that people's intuitive understanding of free will points towards this proposal.

5 Sripada cites his debt to David Shoemaker's excellent paper on caring and agency (2003). Shoemaker writes, "the relation between cares and affective states is extremely tight" (93) and "the emotions we have make us the agents we are" (94).

6 Shepherd and Levy (forthcoming) briefly suggest another idea in this ballpark. They posit that the moral knowledge required to be a responsible agent requires moral perception which requires phenomenal consciousness in order to understand the intrinsic moral value of one's own and others' experiences of pleasure and pain.

3 Emotional Experiences as the Essential Link between Consciousness and Free Will

As we've seen, different theories suggest different connections between free will and consciousness, and the connection might be more or less direct and it might be considered contingent or conceptually necessary. Some accounts (e.g., some libertarian theories) suggest that the connection is direct and conceptual, such that free will, *by definition*, requires consciousness of some sort. More often, the connection, to the extent it is discussed at all, takes a less direct route and suggests a contingent relationship. The idea is that free will requires something *x*, like control, self-awareness, or reasoning, and that *x* is what requires consciousness of some sort, at least given humans' particular cognitive architecture. For example, a reasons-responsive compatibilist might argue that free will requires certain deliberative capacities which happen to require, in some cases, conscious processes in creatures like us. However, on such views, it is unclear whether consciousness, especially phenomenal rather than access consciousness, is necessary for free will or whether it is only contingently related to free will in virtue of the fact that it enables these deliberative capacities that are themselves required for free will. Perhaps, for example, some other complex cognitive agent could carry out the behaviors that are enabled by deliberative capacities *without* phenomenal consciousness.

Again, this suggests that we might be able to test the free will-consciousness connection by considering robots or aliens that are stipulated to have the relevant capacity *x* and behave just like humans but to do so without phenomenal consciousness. To the extent that such creatures are conceivable, we might wonder whether they have free will. If such creatures still plausibly *have* free will, then it suggests a more indirect, contingent relationship between consciousness and free will. However, if the creatures plausibly *lack* free will, even though they behave like humans, it suggests that there must be some more direct relationship, such that the capacity of a creature to be phenomenally conscious is constitutive of or essential for that creature to have free will.

We suggest a relatively direct connection between phenomenal consciousness and free will. The basic idea is that one thing that matters when it comes to being a free agent is that things can really *matter* to the agent. Moreover, in order for anything to matter to an agent, she has to be able to experience the negative and positive consequences of her choices, to be able to feel pain, suffering, and disappointment for bad choices, and to feel pleasure, joy, and satisfaction for good choices, and plausibly to foresee experiencing these feelings when evaluating these choices. Feeling pain and pleasure, and emotions such as anxiety and joy, requires phenomenal consciousness. These mental states

are essential for caring about anything. As Sripada suggests, when someone cares about something,

> Her emotions are bound tightly to the fortunes of the thing… These observations suggest that there is a basic *conceptual* tie between the syndrome of dispositional effects [the functional roles] associated with cares and *what it is* for something to matter to a person. (2016, 1211)

Furthermore, when it comes to consequential or moral decisions involving interpersonal relations, it is essential that the agent can also experience the Strawsonian emotions such as shame, pride, regret, gratitude, and guilt. After all, many of our deepest cares involve other people. So, on this view, the connection between free will and consciousness goes through the capacities to feel emotions that ground mattering, caring, and reactive attitudes.

This view suggests that it is inconceivable for anything to *really* matter to an agent that cannot consciously feel anything, even if that agent were sophisticated and intelligent enough to behave just like us. However, it does seem conceivable that such a robot or alien could *behave* much like us and have the capacities for intelligent action, the evaluation of options, and even complex reasoning, without having phenomenal consciousness. If so, it seems nothing would really matter to such a creature, that it would not really care about what decisions it made. And it seems—to us at least—that it would lack free will.

This, then, is the intuitive connection between consciousness and free will that we wanted to test and compare to other other potential connections, motivated by the pervasive implicit or explicit claims about the consciousness-free will connection, by the relative paucity of explanations for it, and by Shepherd's initial work on this topic.

4 Studies on Attributions of Consciousness and Free Will

Following up on Shepherd's paradigm, we designed two studies to explore people's attributions of consciousness and free will to humanoid robots. The goal was to try to have people consider creatures that look and act like humans while avoiding people's default and implicit attributions of free will and consciousness to humans (and perhaps to any similar biological creatures). As we will see, most people seem to have implicit representations of robots as non-conscious and unfree. In ongoing studies not described here, we use scenarios that describe alien lifeforms.

4.1 *Study 1: Participants and Design*

Our first study had 373 participants (68.2% female, 31.8% male; mean age 19.78, ranging from 18–38), who were undergraduates at Georgia State University. After removing 49 for incomplete data, for missing attention checks, or for spending too little time on the survey, the sample size was 324. This study (as well as its follow-up) was approved by the university's Institutional Review Board, and participation was voluntary and conditioned on informed consent. The study was administered online via Qualtrics.

This experimental vignette study used a between-group design with random assignment to two Learning conditions, as well as to a third control condition. The Learning manipulation varied whether the humanoid robots were able to behave like humans because they could learn based on experiences or because they were preprogrammed with all necessary knowledge. The control condition did not include robots, but instead discussed humans. All scenarios end with a paragraph describing a variety of behaviors that would typically be interpreted as involving conscious experiences in humans, such as feeling cold, expressing empathy, or making a decision. The primary dependent measures consisted of responses to individual statements (from 1 – "Strongly Disagree" to 7 – "Strongly Agree") that were summed together to create sub-scale composite items representing the attributions of the following capacities to either the robots or humans: free will, moral responsibility, basic emotions, Strawsonian emotions, conscious sensation, reason and reflection, and theory of mind.

The experimental vignettes read as follows:

> In the future, humans develop the technology to construct humanoid robots with very sophisticated computers instead of brains and with bodies made out of metal, plastic, and synthetic materials. The robots look, talk, and act just like humans and are able to integrate into human society and to interact with humans across any situation. The only way to tell if they are a humanoid robot instead of a human being is to look inside of them (using x-ray, for example).
>
> (Learning Condition) These robots have various components that process information and allow them to learn from their interactions so that they can change over time. For example, like humans they are able to learn new languages by interacting with people using those languages. Their ability to learn allows the robots to adapt to different situations.
>
> (Pre Programmed Condition) These robots have various components that were pre-programmed with all of the information they would need to behave appropriately in any situation. For example, unlike humans,

> they are pre-programmed to be able to speak any language when interacting with people using the language. Their program allows the robots to behave appropriately across different situations.
>
> Imagine you are asked to observe some of these robots over the course of several weeks, and you see different robots carrying out a wide range of behaviors. For instance, one of the robots, Taylor, gets a hand slammed in a car door and Taylor yells out, grabs the hand, and guards it carefully until it can be fixed. Another robot, Sam, knocks over a glass of water onto Gillian, and apologizes profusely. Another robot, Kelly, comes across a dog whose paw is trapped in a sewer grate and is whining in pain. Kelly gently removes the paw while petting the dog's head. Another one of the robots, Frances, is taking a long walk on a snowy evening, starts shivering, and takes out some gloves and a hat and puts them on. And another robot, Ryan, is at the market purchasing cereal. Ryan stands in the aisle for a minute holding both Corn Flakes and Rice Crispies. Ryan finally puts back the Rice Crispies and places the Corn Flakes in the shopping cart.

In the case of the control vignette, participants were asked to imagine observing some humans acting in the same ways described in the final paragraph. Once participants read one of these vignettes, they were asked to answer questions according to how they understood the issues, not how they thought others might answer. The dependent measures were followed by several manipulation and comprehension checks, and then demographic questions.

4.2 *Study 1: Results*

Prior to the analyses testing our hypotheses, we sought to determine the internal validity of the subscales being used to measure various attributions (see Table 3.1). Following collection of data, coefficients of reliability were calculated for each of our "a priori" subscales. All subscale Chronbach's alpha values were deemed to have an acceptable level of reliability (> .70).

Hypothesis 1: Did attributions of conscious capacities differ between robots and humans, even though they were behaviorally indistinguishable?

Attributions of conscious capacities were subjected to a two-tailed t-test between participants within robot vignettes and those within the human control vignette. There was a significant difference in all measures of conscious capacity attribution, such that participants were less likely to attribute these conscious capacities to robots than they were to humans (see Figure 3.1). More

TABLE 3.1 Subscales and summarized versions of the respective individual statements (roughly one-third of these statements were worded with negations and reverse-scored). Scale validity and reliability was assessed via Chronbach's alpha

Scale	Corresponding questions	Chronbach's Alpha
Free Will (8 items)	These robots have free will, can make choices, have ability to do otherwise, have control over their actions, act of own free will when they act in good and ways we deem (im)moral, are in control (relative to programmers), and what they do is up to them.	.842
Moral Responsibility (5 items)	These robots are morally responsible for their actions, deserve to be blamed (relative to programmers), deserve to be punished for illegal acts, deserve to be blamed for bad acts, and deserve to be rewarded for good acts.	.734
Basic Emotion (9 items)	These robots can feel happiness, frustration, anger, sadness, disappointment, awe, fear, disgust, and can suffer.	.926
Strawsonian Emotion (9 items)	These robots can feel guilt when they do wrong, shame, pride, regret, embarrassment, admiration, indignation, and care about what happens to them and care about what happens to others.	.918
Conscious Sensation (4 items)	These robots *experience,* more than just *process*, the sounds in music, the images in art, the smells of food, and the softness of a blanket.	.913
Reason and Reflection (6 items)	These robots plan, can deliberate, can act for reasons like humans do, can have principles, can reflect on and evaluate their behavior, and can imagine alternative for future actions.	.779

Scale	Corresponding questions	Chronbach's Alpha
Theory of Mind (6 items)	These robots can understand others' emotions, can empathize, can predict what others will do, can infer why others behaved, are aware of their own thoughts, and can understand their own emotional states.	.774

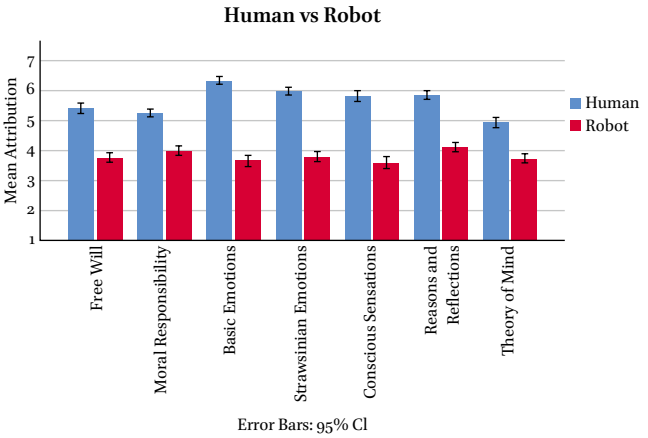

FIGURE 3.1 Robot scenarios evoked significantly less attribution of conscious capacities as compared to the human control scenario. All group differences are significant at $p < .001$ level

specifically, though the robots were behaviorally indistinguishable from humans, participants attributed less free will, $t(322) = -12.62$, $p < .001$, as well as less moral responsibility to them, $t(322) = -10.72$, $p < .001$. Similarly, participants responded that these robots are less able to feel both basic and more complex Strawsonian emotions, $t(322) = -20.00$, $p < .001$, $t(322) = -17.21$, $p < .001$, respectively, as well as less able to experience sensations, $t(322) = -14.73$, $p < .001$. Participants also attributed lower levels of cognitive abilities to the robots, judging them as being less able to reason and reflect, $t(322) = -14.94$, $p < .001$, and less able to utilize theory of mind, $t(322) = -10.28$, $p < .001$.

Note that the attributions of these capacities to robots average near the midpoint, suggesting participants were not very confident about what to say about these robots, which is unsurprising given the minimal information the vignettes provide. However, attributions became more dichotomous once we

examine whether participants are considering the robots to be conscious or non-conscious (see Hypothesis 3 below).

Hypothesis 2: Were participants more likely to attribute conscious capacities to robots that learned as opposed to those that were preprogrammed?

Surprisingly, the ability to learn from experience as compared to being preprogrammed had no discernible effect on any of our dependent measures (all p-values < .10 – data not shown). One possibility is that this information was less important to participants than other information about the robots. Another is that they have implicit representations of robots as fully pre-programmed that are difficult to alter with a few sentences.

Hypothesis 3: Does splitting participants by their response to the question "These robots have conscious experiences" create a divergence in the capacities that they attribute to these robots?

Upon splitting participants based on their response to a dichotomous consciousness question, we found robust differences in the capacities that were attributed to the robots (see Figure 3.2). Those that responded that yes, these robots have conscious experiences, attributed more free will, $t(209) = -5.97$, $p < .001$, as well as more moral responsibility, $t(209) = -4.72$, $p < .001$. As expected, the same results were found for basic and Strawsonian emotions, $t(209) = -7.09$, $p < .001$, and $t(209) = -8.57$, $p < .001$, respectively, as well as the robot's ability to experience sensations, $t(209) = -7.11$, $p < .001$. Similarly, when participants saw these robots as able to have conscious experiences, they also saw them as more able to reason and reflect, $t(209) = -7.70$, $p < .001$, as well as employ theory of mind, $t(209) = -8.01$, $p < .001$. These robust results, though problematized with selection bias,[7] serve to inform the manipulations we developed for Study 2.

4.3 *Study 2: Participants and Design*

Study 2 had 474 participants (69.9% female, 30.1% male; mean age 20.0, ranging from 18–55), who were undergraduates at Georgia State University. After removing 198 for incomplete data, one excluded condition, missed attention checks, or spending too little time on the survey, the final sample size was 278.

7 The grouping and analysis of individuals based on their responses rather than by proper randomization risks an inaccurate representation of the population originally intended to be analyzed.

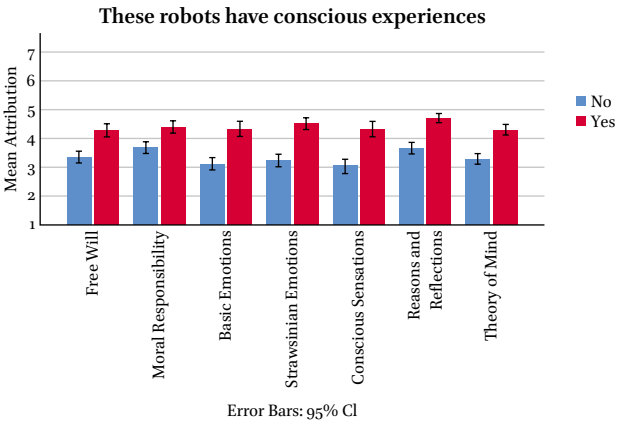

FIGURE 3.2 The dichotomous attribution of consciousness to these robots led to a significant split in all attributions of conscious and other capacities. All group differences are significant at $p < .001$ level

Participation was voluntary and conditioned on informed consent. The study was administered online via Qualtrics.

This follow-up study used a 2 (Learning) x 2 (Consciousness[8]) between-groups factorial design, resulting in random assignment to all possible combinations of conditions (i.e., Learning x Conscious, Learning x Non-Conscious, Preprogrammed x Conscious, and Preprogrammed x Non-Conscious). The Learning manipulation was identical to study 1. The Consciousness manipulation varied whether the robots were described as conscious or non-conscious (as worded below). Participants' responses consisted of responses to individual statements (from 1 – "Strongly Disagree" to 7 – "Strongly Agree") attributing or not attributing different qualities to these robots, that were then combined in order to create the subscales, as described above.

The experimental vignettes were identical to experiment 1 until, following the Learning manipulation, instead of being asked to imagine the robots carrying out various specific behaviors, participants were given further information regarding the robots' mental states and capacities:

> (Conscious) Furthermore, the robots are able to behave just like human beings, and they also have components that enable conscious experiences.

8 For the sake of brevity, a third Epiphenomenal Consciousness condition is not included in this analysis. Responses did not differ significantly from the Consciousness condition, likely because it was difficult to get across the idea of epiphenomenalism.

The robots *actually feel* pain, *see* colors, and *experience* emotions. They do not *just appear* to be conscious when they carry out the same behaviors as humans.

(Non-Conscious) Furthermore, the robots are able to behave just like human beings even though they do not have conscious experiences. They have components that process information such that they can carry out all the same behaviors as humans in just the same ways, but when they do so, they *just appear* to feel pain, *just appear* to see colors, and *just appear* to experience emotions.

The methods and measures were the same as those used in study 1.

4.4 *Study 2: Results*

Hypothesis 1: Did attributions of capacities differ between robots described as conscious versus those that were described as non-conscious?

Attributions of capacities were subjected to a two-way analysis of variance with two Consciousness conditions (Conscious vs. Non-Conscious) and two Learning conditions (Learning vs. Preprogrammed). As predicted, there was a significant main effect of consciousness on free will attributions, $F(1, 274) = 3.89$, $p = .05$, such that participants attributed less free will to robots that were described as non-conscious (see Figure 3.3). Though participants saw conscious robots as more free, they did not judge them to be more morally respon-

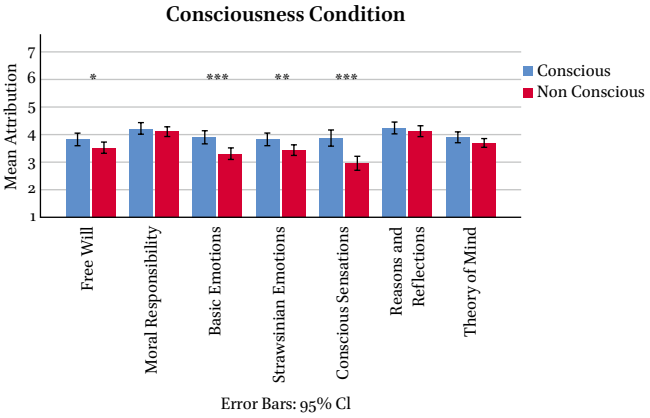

FIGURE 3.3 Compared to robots described as non-conscious, people attributed to robots described as conscious significantly higher Free Will, Basic Emotions, Strawsonian Emotions, and Conscious Sensations. * = $p < .05$, ** = $p < .01$, and *** = $p < .001$

sible for their actions, $F(1, 274) = .80$, $p > .05$. As expected, participants found robots described as non-conscious as less able to feel both basic and more complex Strawsonian emotions, $F(1, 274) = 13.73$, $p < .001$, $F(1, 274) = 6.89$, $p < .01$, respectively, as well as less able to experience sensations, $F(1, 274) = 22.03$, $p < .001$. However, consciousness had no discernible effect on the robots' ability to reason and reflect, $F(1, 274) = .55$, $p > .05$, nor their ability to utilize theory of mind, $F(1, 274) = 2.36$, $p > .05$.

Hypothesis 2: Were participants more likely to attribute conscious capacities to robots that learned as opposed to those that were preprogrammed?

As with study 1, the ability to learn from experience as compared to being preprogrammed had no discernible effect on any of our dependent measures (all p-values < .10 – data not shown).

Hypothesis 3: What attributes (if any) mediated the effect of consciousness on free will attributions?

We used regression analysis in order to investigate potential mediators for the effect of consciousness on free will attributions, and selected mediators based on significant primary paths. In other words, only mediators that were directly affected by the consciousness manipulation were included in the model. Results indicate two primary mediators: participants' attribution of Strawsonian Emotions and Basic Emotions (Figure 3.4). In Step 1 of the mediation model, the regression of the consciousness manipulation on free will attribution, ignoring any mediators, was significant, $b = .31$, $t(276) = 2.00$, $p < .05$. Step 2 showed that the regression of the consciousness manipulation on the mediators Strawsonian Emotions, Basic Emotions, and Conscious Sensations were all significant, $b = .40$, $t(276) = 2.64$, $p < .01$, $b = .59$, $t(276) = 3.72$, $p < .001$, and $b = .91$, $t(276) = 4.71$, $p < .001$, respectively. Step 3 of the mediation analysis showed that, while controlling for the consciousness manipulation, the emotional mediators (Strawsonian and Basic) were each significant predictors of Free Will attribution, $b = .37$, $t(273) = 4.06$, $p < .001$ and $b = .19$, $t(273) = 2.05$, $p < .05$, while Conscious Sensation was not, $b = .04$, $t(273) = .79$, $p = .43$. Step 4 of the mediation analysis revealed that, while controlling for Strawsonian and Basic Emotions (as well the negative control of Consciousness Sensation), the consciousness manipulation was no longer a significant predictor of Free Will attribution, $b = .0034$, $t(273) = .026$, $p = .98$, 95% CI [-.26, .26], indicating full mediation.[9] Thus,

9 The total effect was tested using a bootstrap estimation approach within Andrew Hayes' PROCESS with 5000 samples (2012).

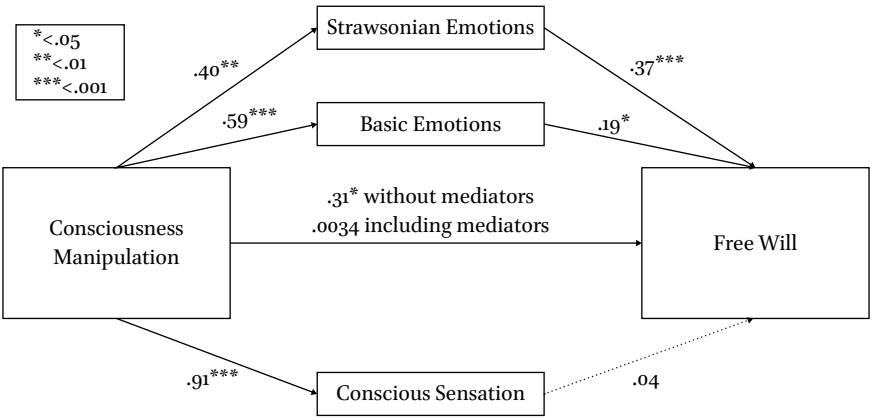

FIGURE 3.4 Analysis of attributes that mediate the relationship between consciousness and free will attributions. Emotional attribution (Strawsonian and Basic) was shown to fully mediate the relationship between consciousness frame and Free Will attribution. Bolded lines indicate significant pathways

it was found that the extent to which people judged the robots as able to experience Strawsonian and Basic Emotions fully mediated the relationship between the consciousness manipulation and people's attributions of Free Will.

These results suggest that phenomenal consciousness plays a particular role in the attribution of Free Will, but not an indiscriminate role. In other words, the ability to feel emotions and have things actually matter to the individual is important in Free Will attributions, yet the ability to have conscious sensations (e.g., the ability to *experience* sounds or smells) specifically plays no significant role.

5 Conclusions

Our results provide some support for our proposed connection between consciousness and free will. However, further studies are clearly required to allow a more fine-grained understanding of which features of consciousness matter most to people's attributions of free will, as well as their relation to attributions of moral responsibility, where we found inconsistent results, and to determine what roles learning and experience play in these attributions, given that our manipulations of whether the robots learn or are fully pre-programmed did not have significant effects. If future research bolsters our initial findings, then it would appear that when people consider whether agents are free and responsible, they are considering whether the agents have capacities to feel emotions more than whether they have conscious sensations or even capacities

to deliberate or reason. It's difficult to know whether people assume that phenomenal consciousness is required for or enhances capacities to deliberate and reason. And of course, we do not deny that cognitive capacities for self-reflection, imagination, and reasoning are crucial for free agency (see, e.g., Nahmias 2018). For instance, once considering agents that are assumed to have phenomenal consciousness, such as humans, it is likely that people's attributions of free will and responsibility decrease in response to information that an agent has severely diminished reasoning capacities. But people seem to have intuitions that support the idea that an essential condition for free will is the capacity to experience conscious emotions. And we find it plausible that these intuitions indicate that people take it to be essential to being a free agent that one can feel the emotions involved in reactive attitudes and in genuinely caring about one's choices and their outcomes. If so, these intuitions support the sort of self-expressive views built on the foundations laid by Strawson and Frankfurt.

We do not want to defend here a metaphilosophical account of the role of ordinary intuitions in philosophical theorizing, a topic of much recent controversy. We will simply conclude by pointing out that most people, along with most theorists, seem to think that consciousness is crucial for free will. Few theorists offer adequate explanations of the connection. If a theorist aims to reject the importance of consciousness to free will, she should explain both what drives most people to think otherwise and why those people are mistaken. If a theorist aims to understand the connection, it might help to understand why ordinary people see it. In fact, we think that understanding why philosophers and non-philosophers alike think that there is a connection between consciousness and free will might suggest strategies for developing plausible theories that explain the connection. If our above results are any indication, these theories will focus on our capacities to *actually care* how others treat us and how we treat them, to feel reactive attitudes in response to such treatment, and to experience the emotions necessary for caring about our decisions and the outcomes of those decisions.

Perhaps, fiction points us towards the truth here. In most fictional portrayals of artificial intelligence and robots (such as *Blade Runner*, *A.I.*, and *Westworld*), viewers tend to think of the robots differently when they are portrayed in a way that suggests they express and feel emotions. No matter how intelligent or complex their behavior, they do not come across as free and autonomous until they seem to *care* about what happens to them (and perhaps others). Often this is portrayed by their showing fear of their own death or others, or expressing love, anger, or joy. Sometimes it is portrayed by the robots' expressing reactive attitudes, such as indignation, or our feeling such attitudes towards them. Perhaps the authors of these works recognize that the robots, and their

stories, become most interesting when they seem to have free will, and people will see them as free when they start to care about what happens to them, when things really matter to them, which results from their experiencing the actual (and potential) outcomes of their actions.

References

Arpaly, N. (2003). *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.

Baumeister, R. (2008). Free will, consciousness, and cultural animals. In J. Baer, J.C. Kaufman, & R.F. Baumeister (eds.), *Are we free? Psychology and free will*, 65–85. Oxford: Oxford University Press.

Bargh, J.A. (2008). Free will is un-natural. In J. Baer, J.C. Kaufman & R.F. Baumeister (eds.), *Are We Free? Psychology and Free Will*, 128–154. Oxford: Oxford University Press.

Berlin, I. (1958). Two concepts of liberty. In *Four Essays on Liberty.* Oxford: Oxford University Press.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18, 227–247.

Buss, S. (2012). Autonomous action: Self-determination in the passive mode. *Ethics*, 122, 647–691.

Caruso, G.D. (2016). Consciousness, free will, and moral responsibility. In R.J. Gennaro (ed.), *The Routledge Handbook of Consciousness*, 78–90. New York: Routledge.

Clarke, R. (2003). *Libertarian Accounts of Free Will*. New York: Oxford University Press.

Dennett, D. (1984). *Elbow Room: The Varieties of Free Will worth Wanting.* Cambridge: MIT Press.

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

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

Frankfurt, H.G. (1987). Identification and wholeheartedness. In F.D. Schoeman (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, 159–176. Cambridge: Cambridge University Press.

Frankfurt, H.G. (1999). *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.

Harris, S. (2012). *Free Will*. New York: Free Press.

Hayes, A.F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://www.afhayes.com/public/process2012.pdf.

Hodgson, D. (2002). Quantum physics, consciousness, and free will. In R. Kane (ed.), *The Oxford Handbook of Free Will*, 57–83. Oxford: Oxford University Press.

Huebner, B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and the Cognitive Sciences*, 9, 133–155.

Jack, A.I., & Robbins, P. (2012). The phenomenal stance revisited. *Review of Philosophy and Psychology*, 3, 383–403.

Kane, R. (1998). *The Significance of Free Will*. Oxford: Oxford University Press.

Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, 7, 67–83.

Levy, N. (2014). *Consciousness and Moral Responsibility*. New York: Oxford University Press.

Libet, B. (1999). Do we have free will? In R. Kane (ed.), *Oxford Handbook of Free Will*, 551–564. New York: Oxford University Press (Reprinted from *Journal of Consciousness Studies*, 1999, 6, 47–57).

Mele, A.R. (2010). Conscious deciding and the science of free will. In R.F. Baumeister, A.R. Mele, & K.D. Vohs (eds.), *Free Will and Consciousness: How Might They Work?*, 43–65. New York: Oxford University Press.

Monroe, A.E., Dillon, K.D., & Malle, B.F. (2014). Bringing free will down to Earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition*, 27, 100–108.

Monroe, A. & Malle, B.F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology*, 1, 211–224.

Nahmias, E. (2018). Free will as a psychological accomplishment. In D. Schmidt, & C. Pavel (eds.), *Oxford Handbook of Freedom*, 492–507. New York: Oxford University Press.

Murray, D. & Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*, 88, 434–467.

Nahmias, E. (2014). Is free will an illusion? Confronting challenges from the modern mind sciences. In W. Sinnott-Armstrong (ed.), *Moral Psychology, vol. 4, Free Will and Moral Responsibility*, 1–25. Cambridge: MIT Press.

Nahmias, E., Morris, S.G., Nadelhoffer, T., & Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research*, 73, 28–53.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*, 41, 663–685.

Penrose, R. (1991). The emperor's new mind. *RSA Journal*, 139, 506–514.

Sartre, J.-P. (1943 [2003]). *Being and Nothingness: An Essay on Phenomenological Ontology*. London: Routledge.

Shepherd, J. (2012). Free will and consciousness: Experimental studies. *Consciousness and Cognition*, 21, 915–927.

Shepherd, J. (2015). Consciousness, free will, and moral responsibility: Taking the folk seriously. *Philosophical Psychology*, 28, 929–946.

Shepherd, J., & Levy, N. (forthcoming) Consciousness and morality. In U. Kriegel (ed.), *The Oxford Handbook of the Philosophy of Consciousness*. Oxford: Oxford University Press.

Sher, G. (2009). *Who knew? Responsibility without Awareness.* Oxford: Oxford University Press.

Shoemaker, D. (2003). Caring, Identification, and Agency. *Ethics*, 114, 88–118.

Smith, A. (2005). Responsibility for attitudes: activity and passivity in mental life. *Ethics*, 115, 236–271.

Sripada, C. (2016). Self-expression: a deep self theory of moral responsibility. *Philosophical Studies*, 173, 1203–1232.

Stapp, H.P. (2001). Quantum theory and the role of mind in nature. *Foundations of Physics*, 31, 1465–1499.

Stillman, T.F., Baumeister, R.F., & Mele, A.R. (2011). Free will in everyday life: Autobiographical accounts of free and unfree actions. *Philosophical Psychology*, 24, 381–394.

Swinburne, R. (2013). *Mind, Brain, and Free Will*. Oxford: Oxford University Press.

Strawson, G. (1986). *Freedom and Belief*. Oxford: Oxford University Press [revised edition 2010].

Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies*, 75, 5–24.

Strawson, P. (1962). Freedom and resentment. *Proceedings of the British Academy*, *48*, 1–25.

Vonasch, A. J., Baumeister, R. F., & Mele, A. R. (2018). Ordinary people think free will is a lack of constraint, not the presence of a soul. Consciousness and cognition, 60, 133–151.

Watson, G. (1975). Free agency. *Journal of Philosophy*, 72, 205–220.

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

Wolf, S. (1990). *Freedom and Reason*. Oxford: Oxford University Press.

Yaffe, G. (2012). The voluntary act requirement. In M. Andrei (ed.), *Routledge Companion to Philosophy of Law*, 174. New York: Routledge.

PART 2

Libet-Style Experiments

CHAPTER 4

Free Will and Neuroscience: Decision Times and the Point of No Return

Alfred Mele

Although I have written many articles and chapters on neuroscientific arguments for the nonexistence of free will, I have not run out of things to say about these arguments. Part of the explanation is that experiments on the topic continue to be conducted and to shed new light on important issues raised by earlier experiments. This is fortunate for me, given that I accepted Bernard Feltz's invitation to write a chapter for this volume.

Experiments performed by neuroscientist Benjamin Libet in the 1980s pose an alleged challenge to the existence of free will. Some neuroscientists have followed Libet's lead, sometimes using electroencephalography (EEG), as he did, and sometimes using functional magnetic resonance imaging (fMRI), depth electrodes, or subdural grid electrodes. In *Effective Intentions* (Mele 2009, chs. 3, 4, and 6), I argued that the neuroscientific work discussed there falls well short of justifying the claim that free will is an illusion. My focus was on the data and on whether the data supported certain empirical claims that have been combined with theoretical claims about free will to yield the conclusion that free will does not exist. There are some interesting new data now. In this chapter, I explore the bearing of some studies published after 2009 on the question whether we have convincing neuroscientific evidence for the nonexistence of free will. Section 1 provides some scientific and terminological background. Section 2 tackles a question about the time at which decisions are made in Libet-style experiments in connection with an examination of a familiar neuroscientific argument for the nonexistence of free will. Section 3 addresses a related neuroscientific argument that features a claim about the point of no return for actions studied in experiments of this kind. Section 4 takes up a skeptical argument that might be thought to have a basis in some recent neuroscientific work. Section 5 wraps things up.

1 Background

Libet makes the following claims:

- 978-90-04-40996-5

> The brain "decides" to initiate or, at least, prepare to initiate [certain actions] before there is any reportable subjective awareness that such a decision has taken place.
>
> LIBET 1985, p. 536

> If the "act now" process is initiated unconsciously, then conscious free will is not doing it.
>
> LIBET 2001, p. 62

> Our overall findings do suggest some fundamental characteristics of the simpler acts that may be applicable to all consciously intended acts and even to responsibility and free will.
>
> LIBET 1985, p. 563

Associated with these claims is the following skeptical argument about free will. Elsewhere (Mele 2018), I dubbed it the *decision-focused skeptical argument*, or *DSA* for short.

1. In Libet-style experiments, all the decisions to act on which data are gathered are made unconsciously.
2. So probably all decisions to act are made unconsciously.
3. A decision is freely made only if it is consciously made.
4. So probably no decisions to act are freely made.[1]

In Mele 2009, I devote a lot of space to showing that premise 1 is not justified by the data and some space to explaining why the generalization in premise 2 is unwarranted. I return to the former matter shortly, after providing some background. An issue with ties to the latter matter is addressed in Section 4.

Decisions to do things, as I conceive of them, are momentary actions of forming an intention to do them. For example, to decide to flex my right wrist now is to perform a (nonovert) action of forming an intention to flex it now (Mele 2003, ch. 9). I believe that Libet understands decisions in the same way. Some of our decisions and intentions are for the nonimmediate future and others are not. I have an intention today to fly to Brussels three days from now, and I have an intention now to click my "save" button now. The former intention is aimed at action three days in the future. The latter intention is about

1 It may be asserted that, strictly speaking, even if no decisions to act are freely made, this does not preclude the existence of free actions. Even if actions that proceed from unfree decisions are not free, what about other actions – actions that do not proceed from decisions (and are not themselves decisions)? I do not pursue this question here. Decisions (or choices) to act are central the philosophical literature on free will, and the discovery that there are no free decisions would be a very serious blow to free will.

what to do now. I call intentions of these kinds, respectively, *distal* and *proximal* intentions (Mele 1992, pp. 143–44, 158, 2009, p. 10), and I make the same distinction in the sphere of decisions to act. Libet studies *proximal* intentions (or decisions or urges) in particular.

In the studies described in this section, participants are asked to report on when they had certain conscious experiences – variously described as experiences of an urge, intention, or decision to do what they did. After they act, they make their reports. The measure of consciousness in these studies is the participants' reports on this matter. As I put it elsewhere (Mele 2009, p. 22), it is "report-level" consciousness that is at issue.

The expression "W time" or "time W" is sometimes used in the literature on Libet's work as a label for the time at which a participant is first conscious or aware of his proximal intention (or decision or urge) to flex and sometimes for the *reported* time of first awareness of consciousness of this. The two times may be different, of course; and Libet himself thought that although the average reported time is about 200 milliseconds (henceforth, ms) before muscle motion begins, the actual average time is about 150 ms before the beginning of muscle motion (1985, pp. 534–35, 2004, p. 128). Here I use "time W" as a label for the actual time of first awareness.

In some of Libet's studies (1985, 2004), participants are asked to flex their right wrist whenever they wish. When they are regularly reminded not to plan their wrist flexes and when they do not afterward say that they did some such planning, an average ramping up of EEG activity (starting 550 ms before muscle motion begins; *–550 ms*, for short) precedes the average reported time of the conscious experience (200 ms before muscle motion begins, –200 ms) by about a third of a second (Libet 1985). Libet claims that decisions about when to flex were made at the earlier of these two times (1985, p. 536).

The initial ramping that I mentioned is the beginning of a readiness potential (RP), which may be understood as "a progressive increase in brain activity prior to intentional actions, normally measured using EEG, and thought to arise from frontal brain areas that prepare actions" (Haggard et al. 2015, p. 325). The significance of RPs is discussed shortly.

Chun Siong Soon and coauthors, commenting on Libet's studies, write: "Because brain activity in the SMA [supplementary motor area] consistently preceded the conscious decision, it has been argued that the brain had already unconsciously made a decision to move even before the subject became aware of it" (2008, p. 543). To gather additional evidence about the proposition at issue, they use fMRI in a study of participants instructed to do the following "when they felt the urge to do so": "decide between one of two buttons, operated by the left and right index fingers, and press it immediately" (p. 543). Soon

and colleagues find that, using readings from two brain regions (one in the frontopolar cortex and the other in the parietal cortex), they are able to "predict" with about 60% accuracy (see Soon et al. 2008, supplementary figure 6, Haynes 2011, p. 93) which button participants will press several seconds in advance of the button press (p. 544).[2]

In another study, Soon et al. ask participants to "decide between left and right responses at an externally determined point in time" (2008, p. 544). They are to make a decision about which of two buttons to press when shown a cue and then execute the decision later, when presented with a "respond" cue (see their supplementary material on "Control fMRI experiment"). Soon et al. report that one interpretation of this study's findings is that "frontopolar cortex was the first cortical stage at which the actual decision was made, whereas precuneus was involved in storage of the decision until it reached awareness" (p. 545).

Itzhak Fried, Roy Mukamel, and Gabriel Kreiman record directly from the brain, using depth electrodes (2011). They report that "A population of SMA neurons is sufficient to predict in single trials the impending decision to move with accuracy greater than 80% already 700 ms prior to subjects' awareness" (p. 548) of their "urge" (p. 558) to press the key. By "700 ms prior to subjects's awareness," Fried et al. mean 700 ms prior to the awareness time that participants later *report*: they recognize that the reports might not be accurate (pp. 552–53, 560). And, unlike Libet, they occasionally seem to treat decisions to press keys as items that are, by definition, conscious (p. 548). Possibly, in their thinking about their findings, they identify the participants' decisions with conscious urges. If that is how they use "decision," their claim here is that on the basis of activity in the SMA they can predict with greater than 80% accuracy what time a participant will report to be the time at which he was first aware of an urge to press 700 ms prior to the reported time. But someone who uses the word "decision" differently may describe the same result as a greater than 80% accuracy rate in detecting decisions 700 ms before the person becomes aware of a decision he already made. These two different ways of describing the result obviously are very different. The former description does not include an assertion about when the decision was made.

There are grounds for doubt about the accuracy of the reported awareness times in these studies. I have discussed such grounds elsewhere (Mele 2009, ch. 6; see also Maoz et al. 2015, pp. 190–94), and I will not do so again here. Instead I focus on two questions. The first question is this: When are the pertinent decisions made in these studies? The second is a question about the point

2 This is not real-time prediction.

of no return in action-producing processes. I introduce it in Section 3, in connection with a skeptical argument that is related to *DSA*.

2 When Do Participants Make Their Decisions?

In Mele 2009, drawing on data of various kinds, I argued that Libet's participants do not make decisions as early as 550 ms before the beginning of muscle motion (–550 ms). Drawing on the same data, I also suggested there that early stages of the readiness potential in his main experiment (a type II RP, which begins at –550 ms) may be associated with a variety of things that are not intentions: "urges to (prepare to) flex soon, brain events suitable for being relatively proximal causal contributors to such urges, motor preparation, and motor imagery, including imagery associated with imagining flexing very soon" (p. 56). Call this group of things *the early group*. As I pointed out, "If RP onset in cases of 'spontaneous' flexing indicates the emergence of a potential cause of a proximal intention to flex, the proximal intention itself may emerge at some point between RP onset and time W, *at* time W, or *after* time W: at time W the agent may be aware only of something – a proximal urge to flex, for example – that has not yet issued in a proximal intention" (p. 57). This point bears on premise 1 of *DSA*, the assertion that in Libet-style experiments, all the decisions to act on which data are gathered are made unconsciously. If proximal decisions to flex – momentary actions of forming proximal intentions to flex – are not made before W, Libet's argument for the claim that they are made unconsciously is undercut.

Also relevant in this connection is evidence about how long it takes for a proximal decision or proximal intention to generate relevant muscle motion. Does it take around 550 ms, as Libet's interpretation of his results implies? I discussed this issue in Mele 2009, where I offered a data-based argument for a negative answer (pp. 60–64). There is additional evidence about this now and about what is represented by readiness potentials.

In Mele 2009, I suggested that some of the participants in Libet's studies may "treat the conscious urge [to flex] as what may be called a *decide signal* – a signal calling for them consciously to decide right then whether to flex right away or to wait a while" (p. 75). Judy Trevena and Jeff Miller later conducted a pair of interesting studies involving a related decide signal. Both studies had an "always-move" and a "sometimes-move" condition (2010, p. 449). In one study, participants in both conditions were presented with either an "L" (indicating a left-handed movement) or an "R" (indicating a right-handed movement) and responded to tones emitted at random intervals. In the

sometimes-move condition, participants were given the following instructions: "At the start of each trial you will see an L or an R, indicating the hand to be used on that trial. However, you should only make a key press about half the time. Please try not to decide in advance what you will do, but when you hear the tone either tap the key with the required hand as quickly as possible, or make no movement at all" (p. 449). (The tone may be viewed as a decide signal calling for a proximal decision about whether to tap or not.) In the always-move condition, participants were always to tap the assigned key as quickly as possible after the tone. Trevena and Miller examined EEG activity for the second preceding the tone and found that mean EEG "amplitudes did not differ among conditions" (p. 450). That is, there were no significant differences among pre-tone EEG amplitudes in the following three conditions: always-move; sometimes-move with movement; sometimes-move without movement. They also found that there was no significant lateralized readiness potential (LRP) before the tone (p. 450). Trevena and Miller reasonably regard these findings as evidence that no part of pre-tone EEG represents a decision to move. The mean time "from the onset of the tone to a key press … was 322 ms in the always-move condition and 355 ms in the sometimes-move condition" (p. 450). If and when the tone was among the causes of a proximal intention to press, the mean time from the onset of that intention to a key press was even shorter. And, of course, muscle motion begins before a key press is complete.

In a second study, Trevena and Miller left it up to participants which hand to move when they heard the decide signal. As in the first study, there was an always-move condition and a sometimes-move condition. Trevena and Miller again found that pre-tone EEG "did not discriminate between" trials with movement and trials without movement, "LRP was absent before the tone," and LRP "was significantly positive after the tone for trials in which a movement was made" (p. 453). They conclude, reasonably, that pre-tone EEG "does not necessarily reflect preparation for movement, and that it may instead simply develop as a consequence of some ongoing attention to or involvement with a task requiring occasional spontaneous movements" (p. 454). Regarding muscle activity, measured using electromyography (EMG), the experimenters report that EMG "seemed to start about 150 ms after the tone" in both "the sometimes-move trials with movements and in the always-move trials" (p. 452). If, in the case of movements, a proximal decision or intention to tap a key followed the tone, then, obviously, the time from the onset of that decision or intention to muscle motion is even shorter. This casts serious doubt on the claim that, on average, proximal decisions or intentions to flex are made or acquired about 550 ms prior to muscle motion in Libet's studies.

As Aaron Schurger, Jacobo Sitt, and Stanislas Dehaene report, "it is widely assumed that the neural decision to move coincides with the onset of the RP" (2012, p. E2909). Like Trevena and Miller and myself, they challenge that assumption. In their view, the brain uses "ongoing spontaneous fluctuations in neural activity" (p. E2904) – neural noise, in short – in solving the problem about when to act in Libet-style studies. A threshold for decision is set, and when such activity crosses it, a decision is made. They contend that most of the RP – all but the last 150 to 200 ms or so (p. E2910) – precedes the decision. In addition to providing evidence for this that comes from the work of other scientists, Schurger et al. offer evidence of their own. They use "a leaky stochastic accumulator to model the neural decision" made about when to move in a Libet-style experiment, and they report that their model "accounts for the behavioral and [EEG] data recorded from human subjects performing the task" (p. E2904). The model also makes a prediction that they confirmed: namely, that when participants are interrupted with a command to move now (press a button at once), short response times will be observed primarily in "trials in which the spontaneous fluctuations happened to be already close to the threshold" when the command (a click) was given (p. E2905).

Short response times to the command clicks are defined as the shortest third of responses to the command and are compared to the longest third (p. E2906). It may be suggested that in the case of the short reaction times, participants were coincidentally already about to press the button when they heard the click. To gather evidence about this, the experimenters instructed participants "to say the word 'coincidence' if the click should ever happen just as they were about to move, or were actually performing the movement" (p. E2907). Participants answered affirmatively in only 4% of the trials, on average; these trials were excluded (p. E2907).

Especially in the case of the study now under discussion, readers unfamiliar with Libet-style experiments may benefit from a short description of my own experience as a participant in such an experiment (see Mele 2009, pp. 34–36). I had just three things to do: watch a Libet clock with a view to keeping track of when I first became aware of something like a proximal urge, decision, or intention to flex; flex whenever I felt like it (many times over the course of the experiment); and report, after each flex, where I believed the hand was on the clock at the moment of first awareness. (I reported this belief by moving a cursor to a point on the clock. The clock was very fast; it made a complete revolution in about 2.5 seconds.) Because I did not experience any proximal urges, decisions, or intentions to flex, I hit on the strategy of saying "now!" silently to myself just before beginning to flex. This is the mental event that I tried to keep track of with the assistance of the clock. I thought of the "now!"

as shorthand for the imperative "flex now!" – something that may be understood as an expression of a proximal decision to flex.

Why did I say "now!" exactly when I did? On any given trial, I had before me a string of equally good moments for a "now!" – saying, and I arbitrarily picked one of the moments. [3] But what led me to pick the moment I picked? The answer offered by Schurger et al. is that random noise crossed a decision threshold then. And they locate the time of the crossing very close to the onset of muscle activity – about 100 ms before it (pp. E2909, E2912). They write: "The reason we do not experience the urge to move as having happened earlier than about 200 ms before movement onset [referring to Libet's partipants' reported W time] is simply because, at that time, the neural decision to move (crossing the decision threshold) has not yet been made" (E2910). If they are right, this is very bad news for Libet. His claim is that, in his experiments, decisions are made well before the average reported W time: –200 ms. (In a Libet-style experiment conducted by Schurger et al., average reported W time is –150 ms [p. E2905].) As I noted, if relevant proximal decisions are not made before W, Libet's argument for the claim that they are made unconsciously fails.

The explanation Schurger and colleagues offer of their findings features neural noise crossing a threshold for decision. Recall Trevena and Miller's suggestion that pre-tone EEG in their experiment may "simply develop as a consequence of some ongoing attention to or involvement with a task requiring occasional spontaneous movements" (2010, p. 454). If Schurger and coauthors are right, this EEG develops partly as a consequence of neural noise, "ongoing spontaneous fluctuations in neural activity" (p. E2904); involvement with the task recruits neural noise as a kind of tie breaker among equally good options.

Given what I have said so far in this section, how plausible is it that Soon et al. found decisions 7 to 10 seconds in advance of a button press? Partly because the encoding accuracy was only 60%, it is rash to conclude that a decision was actually made at this early time (7 to 10 seconds before participants were conscious of a decision).[4] As I observed elsewhere (Mele 2014, pp. 201–02), it is less rash to infer that brain activity at this time made it more probable that, for example, the agent would select the button on the left than the button on the right. The brain activity may indicate that the participant is, at that point, slightly more inclined to press the former button the next time he or she presses. Rather than already having decided to press a particular button next

3 This is not to say that every moment was equally good. I wanted to avoid lengthening my participation in the experiment unnecessarily.

4 Even if the encoding accuracy were much higher, one might reasonably wonder whether what is being detected are decisions or potential causes of subsequent decisions.

time, the person may have a slight unconscious bias toward pressing that button.

What about Fried and colleagues? Did they find early decisions? Their findings are compatible with their having detected at 700 ms before reported W time an item of one of the kinds mentioned in what I called "the early group": urges to (prepare to) press a key soon, brain events suitable for being relatively proximal causal contributors to such urges, motor preparation, and motor imagery. A spontaneous fluctuation in neural activity may be added to the list. If participants made proximal decisions to press a key, the findings are compatible with their having made those decisions as late as the decision time identified by Schurger and coauthors.

3 When Do Participants Reach the Point of No Return?

Some readers who are persuaded that, in the studies I have discussed, decisions are not made at –550 ms or earlier may believe that, even so, the point of no return for the action-generating processes is hit at one or another of the early times identified above and that this is bad news for free will. This belief suggests another argument for skepticism about free will. I call it *PSA*, where *P* stands for "point of no return."

1. In Libet-style experiments, the point of no return for processes that issue in overt actions is reached well before the corresponding decisions are made (anywhere from 550 ms to several seconds before muscle motion).
2. So this is probably true of all decisions to act.
3. If the point of no return for an action-generating process is reached well before the corresponding decision is made, then the decision is not freely made.
4. So probably no decisions to act are freely made.

Someone who endorses premise 3 of this argument may contend that if the point of no return for an action is hit well before the corresponding decision is made, the point of no return for the decision is also hit at this early time and that, in order for a decision to be free, there can be no time prior to the decision's being made at which the point of no return for the making of it has been reached. Not everyone will accept this contention, of course; and not everyone will accept premise 3.[5] But, for my purposes here, premise 3 does not need to be challenged. My concern is premise 1.

5 Compatibilists about free will maintain that free will is compatible with determinism (see McKenna and Coates 2015 for an instructive review). In a deterministic universe, the point of no return for processes is hit very early!

In a comment on the possibility of "vetoing" an urge, intention, or decision, Libet makes the following observation: "In the absence of the muscle's electrical signal when being activated, there was no trigger to initiate the computer's recording of any RP that may have preceded the veto" (2004, p. 141). Given this fact about the design of Libet's main experiment, his data do not allow us look for a point of no return. Suppose we wanted to use Libet's data to test the hypothesis that the point of no return for muscle motion is hit somewhere between 550 and 250 ms before muscle motion starts. What we would like to do is to look at the data to see whether we ever get readings that look like type II RPs during that span of time but are not followed by muscle motion. Libet's experimental design does not allow us to do that. In the main experiment, we get readings only when there is muscle motion.

In Mele 2009, I reported that if I had a neuroscience lab, I would conduct a stop-signal experiment of a certain kind to get evidence about the point of no return in Libet-style scenarios (pp. 75–76). Such an experiment has since been conducted, and it is much more elegant than the one I sketched back then.

The experiment (Schultze-Kraft et al. 2016) takes place in three stages. In all three stages, there is a floor-mounted button and a circle presented on a computer monitor. After the circle turns green, participants wait for "a short, self-paced period of about 2 s" (p. 1080). When they are finished waiting, they may press the button whenever they wish. They earn points if they press while the light is still green, and lose points if they press after the light turns red. The red light is a stop signal.

In stage 1 of the experiment, the stop signals are issued at random times, and the participants are informed of this. Movement times are not predicted. EEG data from this stage are used to "train a classifier to predict upcoming movements in the next two stages" (Schultze-Kraft et al. 2016, p. 1080).

Stage 2 differs from stage 1 in that movement predictions are made in real time using a brain-computer interface (BCI). Participants are not informed of this before the experiment. The aim is to issue stop signals in time to interrupt the participants' movements.

After participants complete stage 2, they are told that the computer had been predicting what they would do and that "they should try and move unpredictably" (Schultze-Kraft et al. 2016, p. 1080). The participants are now ready for stage 3, which is just like stage 2 except that they now have the information I just mentioned.

For my purposes, the most interesting finding is the following one: "Despite the stereotypical shape of the RP and its early onset at around 1,000 ms before EMG activity, several aspects of our data suggest that subjects were able to cancel an upcoming movement until a point of no return that was reached around

200 ms before movement onset" (Schultze-Kraft et al. 2016, p. 1083). If this is when the point of no return is reached, premise 1 of *PSA* is false and the argument crumbles.[6]

4 Another Skeptical Argument

It may be thought that the basis for another skeptical argument about free will is present in my discussion of some recent scientific work. Return to the idea advocated by Schurger and coauthors that the neural decision to move in Libet-style studies is produced when neural noise crosses a threshold. It may be claimed that, if this is what happens, a conscious process is not involved in the (relatively proximal) production of the decision and the decision is therefore unfree. It may also be claimed that this conclusion can properly be generalized to all decisions to act. The argument at issue may be set out as follows. I call it *NSA*, where *N* stands for "noise."

1. In Libet-style experiments, none of the decisions to act on which data are gathered issue from (relatively proximal) conscious processes.
2. So probably no decisions to act issue from (relatively proximal) conscious processes.
3. A decision is freely made only if it issues from a (relatively proximal) conscious process.
4. So probably no decisions to act are freely made.

In Libet-style experiments, participants are instructed to be spontaneous – that is, not to think about what to do. The instructions, if followed, keep consciousness out of the relatively proximal decision-producing process. (Consciousness may be causally involved earlier in the process: for example, participants' conscious understanding of the instructions has a claim to being causally relevant to their performance, and a participant may consciously construct or assent to a plan for completing the experiment, as I did.) The main role for consciousness in these experiments is linked to reporting: participants need a conscious event to report to the experimenter in the case of each particular action. In my case, as I have mentioned, that event was a conscious, silent speech act.

So one difference between decisions made in Libet-style experiments (when participants are following the instructions) and some other decisions is that some other decisions are preceded by conscious reasoning about what to do.

6 This point of no return should not be confused with a later one for the completion of the movement – that is, pressing the button (see Schultze-Kraft et al. 2016, p. 1083).

The existence of this difference challenges the inference made in premise 2 of *NSA*. From the assertion that, when participants are instructed not to think about what to do, consciousness plays no relatively proximal role in the production of their decisions it does not follow that, when people consciously reason about what to do, their conscious reasoning plays no relatively proximal role in the production of their decisions. Someone who wants to argue that, even in the latter case, consciousness plays no role of the kind at issue cannot rely solely on what happens in situations in which there is no conscious reasoning about what to do.

Another difference also merits attention. As I observed, during my stint as a participant in a Libet-style study I arbitrarily picked moments to begin flexing. Arbitrary picking is featured in the Libet-style studies I have described. Participants are said to have decided when to flex a wrist, when to press a key, or which of two buttons to press. There was no reason to prefer a particular moment for beginning to flex a wrist or press a key over nearby moments and (in the study by Soon et al.) no reason to prefer one button over the other. In these experiments, participants select from options they are indifferent about. But in many cases of decision making, we are far from indifferent about some of our options, and many instances of deciding are not instances of arbitrary picking. In typical cases, when we make decisions about matters that are very important to us, after carefully gathering evidence and painstakingly assessing the options, our leading options differ from one another in ways that matter to us, and we do not arbitrarily pick.

The primary upshot of the considerations sketched in this section is easy to see. Arbitrary pickings in Libet-style experiments differ in obvious ways from some of the decisions we make – decisions about matters that we are not indifferent about and that we make after careful reasoning about what to do. And the differences are such that we cannot legitimately generalize from the assertion that conscious reasoning plays no relatively proximal role in the production of the former decisions to the conclusion that this is true of all decisions to act.[7]

5 Conclusion

The primary purpose of this chapter was to bring relatively recent studies to bear on a pair of neuroscientific arguments – *DSA* and *PSA* – for the nonexistence of free will. As I have explained, these studies add to the body of evidence

7 See Mele 2009, pp. 85–87 for a parallel point about premise 2 of *DSA*.

supporting the conclusion that both arguments fail. I also argued that a third skeptical argument, *NSA*, fails as well. I do not want to leave readers with the impression that, in my opinion, the problems with these three skeptical arguments that I have focused on here are the only problems with them. Elsewhere, I have argued against the second premise – a generalizing premise – of *DSA* (Mele 2009, pp. 85–87, Mele 2018), and I have raised worries about the reliability of reported W times (Mele 2009, ch.6).[8]

References

Fried, I., R. Mukamel, and G. Kreiman. (2011). "Internally Generated Preactivation of Single Neurons in Human Medial Frontal Cortex Predicts Volition." *Neuron* 69: 548–62.

Haggard, P., A. Mele, T. O'Connor, and K. Vohs. (2015). "Free Will Lexicon." In A. Mele, ed. *Surrounding Free Will*. New York: Oxford University Press, 319–26.

Haynes, J.D. (2011). "Beyond Libet: Long-term Prediction of Free Choices from Neuroimaging Signals." In W. Sinnott-Armstrong and L. Nadel, eds. *Conscious Will and Responsibility*. Oxford: Oxford University Press, 85–96.

Libet, B. (1985). "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *Behavioral and Brain Sciences* 8: 529–66.

Libet, B. (2001). "Consciousness, Free Action and the Brain." *Journal of Consciousness Studies* 8: 59–65.

Libet, B. (2004). *Mind Time*. Cambridge, MA: Harvard University Press.

Maoz, U., L. Mudrik, R. Rivlin, I. Ross, A. Mamelak, and G. Yaffe. (2015). "On Reporting the Onset of the Intention to Move." In A. Mele, ed. *Surrounding Free Will*. New York: Oxford University Press, 184–202.

McKenna, M. and D.J. Coates. (2015). "Compatibilism." In *Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/compatibilism/.

Mele, A. (1992). *Springs of Action: Understanding Intentional Behavior*. New York: Oxford University Press.

Mele, A. (2003). *Motivation and Agency*. New York: Oxford University Press.

Mele, A. (2003). *Effective Intentions: The Power of Conscius Will*. New York: Oxford University Press.

8 I presented a version of this chapter at the Catholic University of Louvain (August, 2017). I am grateful to the audience for feedback. Parts of this chapter derive from Mele 2009 and 2018. This chapter was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed here are my own and do not necessarily reflect the views of the John Templeton Foundation.

Mele, A. (2014). "Free Will and Substance Dualism: The Real Scientific Threat to Free Will?" In W. Sinnott-Armstrong, ed. *Moral Psychology, Volume 4: Free Will and Moral Responsibility.* Cambridge, MA: MIT Press, 195–207.

Mele, A. (2018). "Free Will and Consciousness." In D. Jacquette, ed. *Bloomsbury Companion to the Philosophy of Consciousness*. London: Bloomsbury Publishing, 371–88.

Schultze-Kraft, M., Birman, D., Rusconi, M., Allefeld, C., Görgen, K., Dähne, S., and Haynes, J.D. (2016). "The Point of No Return in Vetoing Self-Initiated Movements." *Proceedings of the National Academy of Sciences* 113: 1080–85.

Schurger, A., J.D. Sitt, and S. Dehaene. (2012). "An Accumulator Model for Spontaneous Neural Activity Prior to Self-Initiated Movement." *Proceedings of the National Academy of Sciences* 109.42: E2904–13.

Soon, C.S., M. Brass, H.J. Heinze, and J.D. Haynes. (2008). "Unconscious Determinants of Free Decisions in the Human Brain." *Nature Neuroscience* 11: 543–45.

Trevena, J. and J. Miller. (2010). "Brain Preparation Before a Voluntary Action: Evidence Against Unconscious Movement Initiation." *Consciousness and Cognition* 19: 447–56.

CHAPTER 5

Why Libet-Style Experiments Cannot Refute All Forms of Libertarianism

László Bernáth

Introduction

Since Benjamin Libet published the results of his well-known experiments (Libet 1985; Libet–Gleason–Wright–Pearl 1983), it has been heavily debated whether these results refute the existence of free will. Most philosophers who are experts on the topic of free will have reached the conclusion that Libet's original experiments and other Libet-style experiments have not provided enough evidence for denying free will yet.[1] However, the problem as to whether Libet-style experiments could in principle refute free will has not been discussed as much and it seems that there is no consensus on this matter.

Recently, Marcelo Fischborn (2016, 2017) has attempted to shed light on why Libet-style experiments can in principle refute libertarian theories of free will. According to Fischborn, Libet-style experiments can in principle refute libertarian free will because (i) libertarian free will is incompatible with a local determinism in the brain that would make choice predetermined by unconscious brain states and (ii) Libet-style experiments are in principle able to support that there is such a local determinism in the brain.

Against Fischborn, Adina Roskies and Eddy Nahmias (2017) have argued in accordance with their earlier papers (Roskies 2006, Nahmias 2014) that Fischborn is wrong because it is not true either that libertarian free will is incompatible with local determinism or that Libet-style experiments are able to support local or universal determinism.

Although I think that this debate merits attention, both sides share a false presupposition, namely, that the different libertarian theories are similar to each other with regard to what they claim about the role and location of indeterminism in free decisions. Fischborn seems to think that libertarians agree

1 The fact that the philosophical debate about free will is flourishing without discussion of Libet-style experiments proves this point in itself. Nevertheless, there are many works which explicitly state this (e.g. Mele 2009, 2011; Walter 2011, Shields 2014).

- 978-90-04-40996-5

(or should agree) that *decisions about actions*, between which there are no freedom-relevant metaphysical differences, have to be undetermined by previous (mental) states and events in order to the agent has libertarian free will. Roskies and Nahmias argue that only universal (and not local) determinism is incompatible with libertarian free will. Both presuppositions about libertarianism are false and this is why neither Fischborn nor Roskies and Nahmias see that different libertarian theories are vulnerable to the Libet-style experiments to various extents.

In my paper, I spell out which types of libertarian theories can be refuted by Libet-style experiments and which cannot. I claim that, on the one hand, some forms of deliberative libertarianism and restrictive libertarianism cannot even in principle be denied on the basis of these experiments; and on the other hand, standard libertarianism, along with some versions of restrictive and deliberative libertarianism, can in principle be refuted by these experiments. However, any form of restrictive libertarianism can be refuted in the future only if researchers perform new and "untraditional" Libet-style experiments. This is because "traditional" Libet-style experiments investigate decisions in Buridan-type situations. But these decisions are irrelevant with regard to free will, according to the restrictivists.

In the first section, I clarify some terminological issues in order to set the stage for a precise analysis. In the second, I attempt to show what the main reason is for thinking that Libet-style experiments *seem to be* problematic for libertarian theories. Although showing that actions are unconsciously initiated does not pose a real challenge for the elaborated philosophical theories of free will, Libet-style experiments may be problematic for most libertarian theories because, *pace* Roskies and Nahmias, libertarians have good reasons for saying that local determinism and free will are incompatible. In the third section, I argue that because some versions of deliberative libertarianism do not claim that decisions about actions are the center of freedom-relevant indeterminism, Libet-style experiments are unable to refute them after all. In the fourth, I attempt to show why standard centered libertarianism according to which there are no relevant metaphysical differences between different types of conscious decisions is in principle vulnerable to Libet-style experiments to a greater extent. Fifth, I point out the reasons why restrictivist centered libertarianism is less vulnerable to these experiments. Still, I argue that moderate-restrictivism can in principle be denied by Libet-style experiments if these experiments are modified in an important respect. However, proponents of hard-restrictivism should not worry about the possible results of these experiments because these theories restrict the set of free decisions to such a great extent that they cannot be subjected to a proper empirical test.

1 Key Notions of the Debate: "Libertarianism", "Libet-Style Experiments", "In Principle"

In Libet's original experiment (Libet et al. 1983), the subjects are flexed their wrists whenever they wanted. Before they flexed their wrists, they had to memorize the time of their conscious decision to flex when looking at a clock. After flexing their hands, they had to report the time of their initial awareness of their decision or urge to flex their wrists. Libet used Electroencephalography (EEG) to observe electrical activity of the subjects' brains. He found the following. Insofar as the subjects decided spontaneously, the unconscious occurrence of the so-called Readiness Potential (RP) preceded what subject reported to be the time of their initial awareness of their decision or urge by 350 ms. If the subjects did not spontaneously decide, the time gap was even longer. Libet drew the conclusion that the subjects' brain unconsciously initiated their actions, 350 ms earlier than they were aware of their decision. Since, in Libet's view, free will is tied to consciousness, he thought that the subjects did not freely initiate the flexing of their wrists.[2]

Although some philosophers, neuroscientists, and psychologists embrace Libet's conclusions (e.g. Wegner 2002, Hallett 2007),[3] many of them call Libet's conclusions and the adequacy of his whole experimental design into question (Roskies 2011, Walter 2011, and Shields 2014 give useful and brief summaries about the methodological issues). Besides measuring and methodological issues, they worry about whether the electrical signal which was measured by Libet is identifiable with unconscious initiations of actions or whether it is something else. It is not clear even more than thirty years after Libet's first experiments what the RP is besides that it is an electrical activity in the motor cortex that precedes voluntary muscle bursts.[4] Moreover, some call into question whether the measured RP is connected to voluntary action. There is

2 However, Libet claimed that agents have free will because they are able to stop and not perform the initiated action due to so-called "free won't" (Libet 1985).

3 Although Daniel Wegner and Mark Hallett agree with Libet that agents do not freely and consciously initiate their spontaneous actions, they do not believe that 'free won't' or something similar exists, which could be the basis of independent conscious control of actions.

4 The two most influential interpretations of RP are rather different. The first interpretation follows Libet's interpretation according to which RP reflects neural processes associated with unconscious motor preparation of voluntary action. That is, if the RP is earlier than conscious awareness of decision, the initiation of the action is unconscious (Libet 1983, 2004). This interpretation often goes hand in hand with the claim that RP reflects unconscious decision for performing a specific action. The other interpretation claims that RP reflects stochastic fluctuations in neural activity that lead to action following a threshold crossing when humans arbitrarily decide to move (Schurger et al. 2012). In other words, RP reflects such a

some experimental evidence which suggests that the measured RP is a result of Libet's experimental design (Trevena and Miller 2010, Miller et al. 2011).

In order to evade the methodological and measuring issues, many followers of Libet have created more sophisticated versions of the original experiment (Fried et al. 2011). Others, most notably Chun Siong Soon and his colleagues (Soon et al. 2008, 2013), have made more radical changes besides fixing these measuring and methodological problems. Soon and his colleagues did not use EEG. Instead, they focused on observing specific brain regions using fMRI. Since the RP is defined in such a way that it can be detected only using EEG, using fMRI means that Soon's experiment does not attempt to find a correlation between RP and the conscious initiation of action. Moreover, they investigated not only choices about bodily motions but about mental actions as well, given that their later experiment examined decisions about adding or subtracting numbers. They found that the outcomes of choosing between pushing a right-sided or a left-sided button were predictable on the basis of patterns of activation in specific brain regions with roughly 60% accuracy, and 7 seconds before the conscious experience of the choice (Soon 2008). They had similar results in the other experiment with regard to abstract spontaneous choices apart from that they could predict with 60% accuracy only 4 seconds before the conscious awareness of choices (Soon 2013).

Even though Libet's early experiments and Soon's experimental design are very different, both of them are called Libet-style experiments. Why should we put them into the same category? Because I do not think that one should call all neuropsychological experiments about free will Libet-style experiments, and since I did not find any definition of what makes an experiment to be a Libet-style experiment, I propose the following list of the necessary (and together sufficient) conditions for being a Libet-style experiment.

i. By means of neuroscientific methods, Libet-style experiments try to identify such unconscious brain activities/states that determine (or which are identical with) decisions about actions through identifying correlations between brain activities/states and outcomes of decisions.
ii. Libet-style experiments examine decisions that are considered to be free decisions by the majority of laymen and/or philosophers.
iii. In order to ensure that the experiment will investigate decisions that are considered free, the experiment examines decisions in decision-situations

neural mechanism that helps us to make decisions in situations in which there is no good reason for choosing one alternative over another.

for which the influence of unknown intentions and preferences can reasonably be excluded.

iv. During Libet-style experiments, subjects have to repeat the same type of decision many times.

v. The researchers have to identify the time when the subjects were aware of making their decision.

Although most of these conditions on being a Libet-style experiment are clear without further explanation, it is worthwhile focusing on (iii.) a little bit more. The reason why the influence of unidentified or unknown prior intentions and preferences should be excluded somehow is that, in many cases, prior conscious states and intentions explain the exact outcome of the decision. Let us suppose that someone consciously decided that whenever she has to decide between two identical buttons, she will push the one on the left. After a while, she faces this "dilemma", and she instantly pushes the left button without conscious deliberation. Even though this freely-formed intention was the main cause of her action, pushing the button was such a fast reaction that it initiated action unconsciously. Even if a neuroscientist could detect that one of her brain states which was unconscious at the time of the initiation of action determined whether she pushed the right or the left button, many laymen, compatibilist, and libertarian philosophers would think that it does not prove that her decision was not free, because it is probable that the unconscious brain state in question is identical with (or is the neural basis of) her previously freely-formed intention to press the left button. In order to evade these kinds of interpretative possibilities, the researchers have to choose such decision-situations in which they are able to exclude the influence of such distal (and free) preferences or intentions. Without focusing their attention on such a decision-situation, they cannot be sure whether, in some sense, a prior consciously formed intention initiated the action instead of an unconscious one.

This is the reason why Libet and most of his followers examine decision-situations in which there are no objective reasons to prefer one alternative over another one. In Libet's original experiments, the subjects have to decide when to flex their wrists without any good reason to flex their wrists at any particular time during the whole experiment. In Soon's experiments, the subjects have no objective reason to prefer pushing the left button/adding over pushing the right button/subtracting.

I call these experiments traditional Libet-style experiments and I will claim that these cannot refute many libertarian theories. The reason is the following. Although restricting the focus of the experiment in such a way is a good way

to exclude the influence of prior and unknown intentions and preferences, it has the disadvantage that the researcher can examine only one type of decision. These are Buridan-type decisions in which there is no reason to choose one alternative instead of any other. But there are some good reasons to think that these Buridan-type decisions are so different to other types of decisions that one cannot generalize results about the former decisions to the latter ones.

As I will argue, this issue has great importance with regard to the refutation of libertarian theories. Luckily, there are other possible ways to exclude the influence of unknown and prior intentions and preferences. For example, Maoz and his colleagues (Maoz 2017) asked their subjects about their relevant preferences in order to reasonably exclude the possibility that there are unknown relevant preferences influencing the choices of the subjects during the experiment. Maoz's research is a new form of Libet-style experiments, and I consider it as one of the first instances of an untraditional Libet-style experiment.

The second terminological issue that I should handle is what it means for an experiment to be capable "in principle" of supporting or proving a particular claim. Fischborn has an illuminating example.

> It can be true that physics could in principle show that time travel is possible even if time travel is actually physically impossible. Saying that physics can in principle show that time travel is possible (if time travel is actually possible) says nothing about the truth or plausibility of the possibility of time travel; it only says that physics is the right science for an investigation on the possibility of time travel.
>
> FISCHBORN 2017, 2

One can paraphrase this quote regarding Libet-stlye experiments in the following way. Saying that Libet-style experiments can in principle show that free will does not exist (if free will actually does not exist) says nothing about the truth or plausibility of the inexistence of free will; it only says that Libet-style experiments are the right experiments for an investigation on the (in)existence of free will.

Fischborn's example is extremely useful, because it can help shed light on an important issue. In general, if one asks whether time travel is possible, she means about "time travel", among other things, that *the body* of the traveler has to travel through time in order that the traveler counts as a genuine time traveler. In this case, physics is the right science for an investigation on the possibility of time travel. However, in an unusual case, one can mean by time travel

such a process in which only a body-independent soul travels to the future or the past. In this unusual case, physics is not the right kind of science for an investigation on the possibility of time travel.

Similarly, it depends on what one means by the term 'free will', if we want to know whether Libet-style experiments can in principle support claims that refute its existence. If one accepts a Hobbesian notion of free will according to which one has free will if she is able to act in accordance with her desires, then Libet-style experiments cannot even in principle show that one has no free will. This is because Libet-style experiments can primarily support only such claims which are based on the "correlation" of some unconscious brain states/events and conscious outcomes of decisions. But these correlations do not say anything about whether the agent has a capacity thanks to which she can act in accordance with her desires. (Because even if such brain states determine our decisions, it does not mean that our actions and desires do not fit each other in most of the cases).

Here, my main interest is whether Libet-style experiments are able to refute libertarian theories of free will. Libertarianism is the conjunction of two theses. One is the thesis of incompatibilism according to which universal determinism is incompatible with free will. The second is the free will thesis which says that adult humans without severe psychological disorders have free will. Even though all libertarian theories accept these theses, there are great differences between them. This is partly because they disagree about i) the location of the relevant indeterminism and ii) what the exact metaphysical structures of different types of decisions are.

I claim that, contrary to Fischborn, Nahmias, and Roskies, the differences which are rooted in these disagreements are relevant with regard to whether libertarian theories can be refuted by Libet-style experiments. Traditional forms of Libet-style experiments can in principle refute only those which claim that a) the location of freedom-relevant indeterminism can be found in decisions about actions *and* b) there is no freedom-relevant difference between Buridan-type decisions and other non-coerced decisions. Consequently, they can refute standard libertarianism *and* some deliberative libertarian theories. Untraditional Libet-style experiments, if they are performed, will be able to refute more types of libertarianism. In principle, they can threaten those which claim that α) the location of relevant indeterminism can be found in the decision process about actions β) there is no freedom-relevant difference between hard decisions in which the agent is strongly motivated to perform more than one action. But neither traditional nor untraditional Libet-style experiments can refute those libertarian theories which say either that A) the location of freedom-relevant indeterminism is not necessarily in the decision processes

about actions or that B) there is a freedom-relevant difference between decisions under moral temptation and every other type of decisions. In the following sections, I will argue for these claims in detail.

2 How Are Libet-Style Experiments Able to Get Libertarians into Trouble?

In the literature, there are two models about how Libet-style experiments deny the existence of free will (libertarian or not). According to the first, Libet-style experiments are relevant to the free will debate because they support the claim that we do not consciously initiate our actions. If it is true, then, as the argument goes, it is also true that our conscious considerations do not influence our actions. And this would mean that we do not have free will. (This kind of argumentation, most prominently, can be found in Wegner 2002. Daniel Wegner argues that even intentions have no causal role in producing our actions).

In my view, this argument is problematic for many reasons. For the sake of brevity, I would like to stress only one difficulty (more detailed criticism can be found in Mele 2009). Even if it is true that we do not consciously initiate our actions, this does not necessarily mean that our conscious considerations do not influence our actions. If a soccer player consciously reaches the conclusion that she should try to deflect the ball after the free kick, her conscious decision influences her bodily motion even if, later, she unconsciously initiates her bodily movement which is aimed at deflecting the motion of the ball. In light of this, even if it is an illusion that there is such an activity as conscious initiation of action, it is still quite possible that our conscious considerations have a major role in controlling our actions.

This problem merits attention because the vast majority of free will theories do not consider the conscious initiation of action as a necessary condition on free will and moral responsibility.[5] Rather, they stress the importance of specific causal or non-causal influence of the agent and her reasons for action. This is why, if we resist the problematic conclusion that Libet-style experiments can show that conscious considerations do not influence our actions, this strategy seems to be harmless after all.

5 Timothy O'Connor (2000, 122) has lamented that elaborated philosophical theories of free will do not clarify the role of consciousness with regard to free actions. As far as I can tell, experts on this topic still tend to neglect this issue. However, there are some exceptions, for instance, Hodgson 2012.

Now, one may object that Libet-style experiments can prove not only that we unconsciously initiate our actions but also that we unconsciously decide what we will do. In traditional Libet-style experiments, the subjects have no reason to prefer one alternative over the others. Thus, the initiation of action cannot be influenced by conscious reasons and previous conscious decisions. Therefore, the unconscious initiation of the action is the very source of the action. Since the very sources of our actions are the decisions about actions, one can say that the unconscious initiation of the action is the decision itself.

There are two problems. First, even if the unconscious initiation is the main source of the action in question, it does not necessarily mean that considering this initiation to be a decision is a good idea. Let us suppose that our universe is deterministic and the Big Bang is the main source of every event of the universe. This surely does not mean that the Big Bang is any kind of decision. One may argue that there is a good reason for considering unconscious initiations to be decisions from an action-theoretical point of view. But, as far as I know, there is no such argument yet. Second, even if the unconscious initiation is the main source of the action in Buridan-type situations, it does not necessarily mean that it is the case even in situations in which the agent has reasons for choosing one of the alternatives. It is still possible that conscious reasons causally influence this initiation even if the initiation is unconscious. And in this case, most theories of free will is not get into trouble.

The second model for how Libet-style experiments can in principle refute theories of free will is spelled out by Marcel Fischborn. According to Fischborn, who focuses on libertarian theories, Libet-style experiments are in principle able to refute the existence of free will because they can support the following kind of laws.

> LD1. For any event *x*, and any subject *s*, if an *x* that is a pattern of neural activity of type *B* occurs in *s*'s brain, then *s* will decide to push a given button.
>
> FISCHBORN 2016, 497

Fischborn suggests a particular way of supporting LD1. Let us suppose that Soon and colleagues will find in the future that there is a 100% correlation between some types of unconscious brain states and some outcomes of Buridan-type decisions. In this case, the neuroscientists could reasonably conclude that there are LD1-like psychological laws which imply local determinism.

Moreover, Fischborn claims that if the neuroscientists could prove many instances of LD1, they would be reasonably able to generalize their results.

Thus, they could in principle support the claim that not only LD1-type laws are true but the following one as well.

> DNC. For any subject *s*, any choice *x*, and any course of action *X*, if *s* chooses to do *X*, then there is a previous event *y* of a type *Y* in *s*'s brain, such that whenever an event of type *Y* occurs in someone's brain, then this subject will choose for the course of action *X*.
>
> FISCHBORN 2016, 497

But if DNC was true, it would mean that our decisions are locally determined by unconscious brain states. From this, according to Fischburn, it follows that libertarian theories of free will are refuted because they are incompatible with this kind of local determinism.

Adina Roskies and Eddy Nahmias have argued that this way of refuting libertarian theories does not work (Roskies & Nahmias 2017). To begin with, it is unsure that any Libet-style experiments (or other neuroscientific experiments) will support anything similar to LD1. This is because it may be the case that there are multiple realizations of mental states, and/or the thesis of extended cognition may be true, and/or the claim that neural activities are complex and chaotic might be true. Insofar as one of these views is on the right track, it is very unlikely that neuroscientific experiments will find laws similar to LD1, because they imply that the relations between brain states and mental states are too complicated for the existence of universal and deterministic law-like relations between brain states and decisions.

To my mind, Fischborn has answered to this argument in a very plausible way (Fischborn 2017, 199–200). He has pointed out that this objection has nothing to do with the question whether Libet-style experiments can *in principle* prove LD1. For the sake of argument, Fischborn accepts that if any of the aforementioned hypotheses are true, then Libet-style experiments cannot support LD1. (I would even say that this is because if one of these suppositions is true, then LD1 cannot be true). But it does not mean that Libet-style experiments cannot *in principle* support LD1. If there are psychological laws such as LD1, Libet-style experiments could support this claim partly because, if LD1 is true, then neither multiple realizability, extended cognition, nor the thesis that neural networks are complex and chaotic are true. In other words, if LD1 is true, Libet-style experiments can show that it is the case.

Roskies' and Nahmias' second objection is based on the claim that libertarian free will is incompatible only with universal determinism. Thus, even if DNC could be proven, it would not refute libertarianism. Roskies and Nahmias are aware that some libertarians have explicitly claimed that free will *is*

incompatible with local determinism (they mention Kane 1996, Ekstrom 2000, Balaguer 2012). However, Roskies and Nahmias argue that libertarians do not have good reason for worrying about local determinism because all of their arguments for incompatibilism support only the incompatibility of universal determinism and free will.

In my view, libertarians argue mainly against the compatibility of *universal* determinism and free will because they try to deny all forms of compatibilism regardless of their exact details. But *most* libertarians deny that agents may have free will if all of their decisions are determined by unconscious brain states. Even if they do not explicitly claim this, one can be sure about it because most libertarians think that the *decisions about actions* should be undetermined by preceding brain or/and mental states in order to count as (directly) free.[6] More precisely, most libertarians claim that the very event that is choosing an action has to be undetermined. In the literature, these libertarians are called centered libertarians, and most of the detailed libertarian theories fall into this group.[7]

Moreover, centered libertarians do not only claim that *momentary free decisions about actions* are the indeterministic sources of free will but they have *good reasons for* embracing this view (*pace* Roskies & Nahmias 2017). In order to grasp why it is the case, one has to see what libertarians should do for dialectical reasons besides refuting compatibilism. First, they have to show how satisfying the libertarian control-conditions of free action provides more control over action than satisfying compatibilist control-conditions. This is because libertarians, by definition, deny that satisfying compatibilist conditions of free action is sufficient for free action. And if universal determinism rules out free action, the reason for it has to be that it undermines the satisfaction of the control condition on free actions. (The epistemic condition, which is the other widely accepted condition for being morally responsible, has nothing to do with the truth or falsity of determinism.) Second, they need to explain also how the outcome of an indeterministic process need not be a matter of luck. They have to solve this problem because if an action is simply a matter of luck, it cannot be controlled by anything. And an uncontrolled action cannot be free (Levy 2011). Third, they should provide some evidence for that such an indeterministic process which is not simply a matter of luck exists.

6 That is, the location of indeterminism has to be found somewhere in the decision-process. As far as I know, there is only one elaborated libertarian theory which does not accept this claim (Ekstrom 2000). In the next section, I will investigate this theory.

7 According to the article about libertarian theories of free will at Stanford Encyclopedia (Clarke & Capes 2017), all agent-causal and non-causal theories are centered accounts. Even most event-causal libertarians have elaborated centered theories.

In order to solve the second issue, the libertarians have to posit not only an indeterministic process but a *controlled* indeterministic process. In order to reply to the first issue, they have to suppose a controlled indeterministic process that is controlled differently with respect to the way that the compatibilist agent controls her action. Most libertarians have reached the conclusion that the moment of deciding is the event which has to be undetermined in order to solve these problems because decisions about actions are considered to be those events that are controlled in the most direct way. So there is a *hope* that if decisions about actions are the key indeterministic events which are not determined by previous brain/mental states, then they are still *highly controlled* indeterministic events. Additionally, some libertarians think that there is good evidence for libertarian free will if the moment of deciding is considered as the key indeterministic free event. They argue that the phenomenological characteristics of these decisions can serve as an evidence for the existence of highly controlled but undetermined events. At the moment of the decision, the choice seems to be both undetermined and controlled from a first person point of view,[8] and – according to these libertarians – this phenomenological fact is a proper evidence for believing in the existence of sufficiently controlled and undetermined decisions until there are no strong counter-evidences that outweigh this phenomenological consideration.[9]

Thus, from a libertarian perspective, there are good reasons for denying even psychological and local determinism because they have good reasons to claim that the very moment of decisions about actions are those indeterministic and highly controlled events that are the sources of free will. Or at least, these reasons are no worse than other philosophical reasons for embracing other philosophical views. Rejecting the relevance of local determinism with regard to libertarianism without providing arguments against these reasons is insufficient.

Before I address the problem as to whether Libet-style experiments are in principle able to refute centered libertarian theories, I investigate which non-centered or deliberative libertarian theories are vulnerable to Libet-style experiments.

8 However, some doubt that these decisions seems to be undeterminded from a first person perspective. See Nahmias et al. 2004, Horgan 2011, 2012.

9 This phenomenological argument for free will can be found in Reid 1788: 36; C.A. Campbell 1957: 168–174; O'Connor 1995: 196–197; Swinburne 2011: 82. Deery 2015 reconstructs the argument with great care, and I borrowed his list about where the argument explicitly appeared.

3 Are Libet-Style Experiments Able to Refute Deliberative Libertarianism?

In the literature, those libertarian theories are called deliberative which claim that the location of freedom-relevant indeterminism is not (or at least not necessarily) at the *moment* of choice. For example, Laura Ekstrom (2000) claims that it is enough for having free will if the agent's preference formation is indeterministic. Surprisingly, as far as I know, Ekstrom is the only devoted libertarian who adopts a deliberative libertarian theory. However, some experts who officially do not endorse libertarianism have outlined other deliberative libertarian approaches. Daniel Dennett (1978) and Alfred Mele (1995 ch.12) suggested two very similar theories for libertarians. According to Dennett, libertarians should claim that the processes determining which considerations occur to one are the heart of libertarian freedom. Similarly, Mele (1995, 2006: 9–14) shows that the libertarian may claim that the key indeterministic processes are those which determine whether a belief (or desire) comes to the agent's mind.

In principle, Mele's and Dennett's suggestion could be refuted by Libet-style experiments. Let us suppose that someone performs an experiment which is very similar to Soon's and his colleagues' experiment. And say she finds that there is a 100% correlation between different patterns of unconscious brain activities which precede the decision by 10 seconds and the different possible outcomes of the decision. Since it is plausible to suppose that the processes that determine which considerations occur to one are not finished in Buridan-type situations 10 seconds before the choice, this experiment would support strongly the claim that these processes are determined, as well as the outcomes of choices, by unconscious brain states.

However, it is not so simple in the case of Ekstrom's deliberative libertarianist view. Ekstrom's claims, in contrast with other libertarian approaches, that not only decisions about actions but also decisions concerning what desirable is can be undetermined and metaphysically free. According to Ekstrom, the latter kind of decision are not necessarily about what would be desirable to do at a given time but can be about what the good is in general (Ekstrom 2000, 106–109).

Prima facie, this version of libertarianism is not as vulnerable to Libet-style experiments since these experiments investigate only decisions about actions. Even if Libet-style experiments show that decisions about actions are determined by unconscious brain states, it is open to Ekstrom to claim that decisions about what is good are relevantly different in this regard because they are undetermined and free. She could argue, for instance, that all decisions about

actions are driven by conscious or unconscious brain states because our most valuable and rational choices are based on what we consider as good or desirable. But decisions about what is good and what kind of things are desirable may not determined by other choices. This kind of defense of free will cannot be refuted by reference only to Libet-style experiments because they investigate only decisions about actions. One should deny this argumentation on the basis of philosophical insights or empirical data. But the neuropsychological experiments which could test whether decisions about what is good are determined by previous brain states should be so different to Libet-style experiments that it would be misleading to call them "Libet-style experiments".

If am right, in contrast with Dennett's and Mele's deliberative libertarian theories, Ekstrom's view, even in principle, cannot be refuted by Libet-style experiments. To put it simply, this is because Libet-style experiments do not investigate decisions about what is good (this may be considered to be judgment rather than decision by many philosophers). To generalize the moral of this, Libet-style experiments are not able to deny those libertarian theories according to which it is not (or not only) decisions about actions that are the location of free will.

4 Centered Libertarianism

4.1 *Standard Versions*

Most libertarians claim that decisions as momentary choices about actions should be undetermined to count as (directly) free. The standard version of this approach says that there is no relevant difference between non-pathological decisions with regard to metaphysical freedom. Every decision is directly free, there are no non-pathological decisions which are unfree or only indirectly free because they are the consequences of prior directly free decisions. In other words, every psychologically "normal" decision is *directly* free regardless of the particular type of the decision in question. This version of libertarianism is the most common. All non-causal (Goetz 1988, Ginet 1990, McCann 1998, Lowe 2008, Pink 2004) and agent-causal theories (Chisholm 1966, Taylor 1992, O'Connor 2000, Clarke 2003, Griffith 2007, Steward 2012) fall into this group. (I know about only one exception. It is Swinburne 2011, which defends a restrictive agent-causal theory). Even some event-causal accounts endorse this unrestricted version of libertarianism (Nozick 1981, Hodgson 2012).[10] In this section, I would like to point out that this unrestrictive view on free decisions

10 There are accounts which are between restrictivist and non-restrictivist views. They accept that there are indirectly free and determined decisions but they do not say that

has the price that these standard centered libertarian theories are quite vulnerable to Libet-style experiments.

Libet-style experiments can, in principle, show that decisions are predetermined in Buridan-type situations. For instance, if Soon and colleagues were able to predict on the basis of the activity of unconscious brain states whether the agent will push the left or the right button with 100% accuracy 7 seconds before the conscious awareness of choice, then this would provide strong evidence for unconscious determination with regard to Buridan-type decisions. But, and this is the point, insofar as there is no relevant metaphysical difference between non-pathological conscious choices, this result would provide strong evidence that all conscious decisions are determined by unconscious brain states. Without question, if it turned out to be true, it would refute centered libertarianism.

But why do so many libertarians think that there is no relevant metaphysical difference between non-pathological conscious decisions? The reason why is that there is no *phenomenological* difference between conscious decisions with regard to the prior indeterminacy of these decisions. All conscious decisions about actions seem to be undetermined from a first person point of view. At the moment of conscious decision, agents have the impression whether they decide for *A* alternative or *B* alternative is settled only at the very moment of the choice in question. This phenomenological characteristic of decisions is relevant for many libertarians because they regard it as the best evidence for having free (highly controlled but indeterministic) decisions. If libertarians do not claim that all decisions which phenomenologically seem to be free actually are free, then they cannot maintain that phenomenological traits reliably indicate whether a decision is actually free; therefore, they have to look after another justification for believing in free will. And this is not an easy task. Consequently, even if claiming that every non-pathological decision is free has the cost that the theory will *in principle* be vulnerable to Libet-style experiments, it may be worthwhile sticking to this idea. Until standard centered libertarianism is only *in principle* vulnerable to Libet-style experiments and not refuted by them, the proponents of standard centered libertarianism can reasonably hope that their theory will not be refuted even in the future.

However, many who initially regard centered libertarianism as a plausible view on free will would not like to take this risk. After all, this conclusion about standard centered views says that even a version of the well-known Libet-style

specific types of decisions are necessarily determined. For instance, it seems to me that Mele (2006) offers such a libertarian account.

experiments can *in principle* show that free will is only a *phenomenological illusion*. Luckily, for those who would like to neither reject centered libertarianism nor take this risk, there are other versions of centered views that are *in principle* less vulnerable to Libet-style experiments.

4.2 *Restrictivist Versions*

Libertarian restrictivists claim that only some types of our decisions about actions are *directly* free (Campbell 1938/2013, van Inwagen 1989, Balaguer 2004; Kane 1996, 2000, 2007). This is because, in their view, only some types of decisions can be indeterministic and sufficiently controlled at the same time. However, they tend to call "free" even determined or not sufficiently controlled decisions which are properly influenced by *directly* free decisions.

There are, so to say, hard-restrictivists who hold that only one type of decision can be directly free. For instance, C.A. Campbell (1938/2013) famously claimed that only decisions that are made in situations of moral temptation are directly free. Other restrictivists, for instance Peter van Inwagen and Robert Kane, say that there are more types of decisions which are directly free. Kane claims that all decisions are directly free in which either the agent has more than one motivationally viable option or the agent must make efforts to sustain her intentions against desires and other conditions which make it difficult to carry out our purposes once chosen (Kane 1996, 2000, 2007). I call this more permissive kind of view moderate-restrictivism.

These theories agree that decisions are not free in Buridan-type situations. This is because, in Buridan-type situations, agents have no reasons to prefer one possible course of action over another (or over doing nothing at all). For example, in Libet's original experiments, the agents do not have any reason for prefer flicking their wrists over doing nothing at any time during the experiment. The subjects are not motivated in either way; thus, neither option is motivationally viable for them. At best, only one option is motivationally viable from trial to trial from their perspective, and it is that which is suggested for them by their unconscious processes. So showing that those Buridan-type decisions in which the agent has no reasons to prefer any possible course of action are not free is not in contradiction with restrictivist libertarianism. Consequently, results of traditional Libet-style experiments which investigate Buridan-type decisions are, even in principle, not able to refute libertarian restrictivism.

One may disagree on the basis that restrictive libertarians believe that Buridan-type decisions are indeterministic (van Inwagen 1989). So, the objection could go, even traditional Libet-style experiments are able to refute the views of restrictivist libertarians about decisions.

It is true that traditional Libet-style experiments can in principle refute the *views of restrictivists about Buridan-type decisions* but it does not mean that these experiments are able to refute libertarian restrictivism. For what libertarian restrictivism argues is that Buridan-type decisions are not free. Libertarian restrictivists can justify this claim by saying either that they are not free because the outcome of these decisions are a matter of chance or that they are not free because they are determined. Peter van Inwagen supposes that Buridan-type decisions are based on an "internal coin-tossing", suggesting that the outcome of these decisions are just as lucky as the outcome of a coin-tossing (van Inwagen 1989, 417). However, in the light of traditional Libet-style experiments, the restrictivist could suggest that RP is neural noise that helps to form intentions through generating an urge to choose something in such cases that the agent has no reason to choose any particular alternative. It is even open for the restrictivist to say that neural noise is determined by the exact state of the brain and that it in turn determines which alternative will be chosen. In this case, the restrictivist should not worry even if the researchers could predict the outcome of the decision with 100% accuracy in Buridan-type situations. The restrictivist could say that the neural noise in question has a deterministic role only in Buridan-type situations because, in other situations, the agent decides on the basis of her reasons.

Although traditional Libet-style experiments are harmless against restrictivism, untraditional Libet-style experiments may get moderate-restrictivism into trouble. Moderate-restrictivists agree that if the agent has to decide between two options which, for the agent, seem to be attractive on the basis of different sets of reasons, she will make a directly free decision. For instance, if one has to decide between donating towards the education of poor children and donating towards the care of sick people, and if one has not previously ranked the sets of reasons which motivate these actions, then the one will make a directly free decision.

Even though this decision could be not be investigated by traditional Libet-style experiments, Maoz and his colleagues performed a Libet-style experiment which could, in principle, show that even these hard decisions are determined (Maoz 2017). First, they instructed the subjects to rate how much they would like to support different non-profit organizations (NPO) with a $1000 donation on a scale of 1 to 7. After that, the subjects had to decide between different NPOs. In some cases, they had to choose while knowing that their choice would raise the probability that the researcher would in fact give a donation for the chosen NPO. In other cases, they know that their choices will not influence whether the chosen or another NPO will get donations. Moreover, the researchers asked the subjects to make completely arbitrary decisions in these latter cases. The

researchers investigated, among other things, the RP signals during the deliberate and arbitrary decisions.

This experimental setup, in principle, could support the claim that even those decisions that are both difficult and based on reasons are determined by unconscious brain states. Let us suppose that they find type-A RP signs if the subject chooses the NPO which is represented at the left side of the screen, and type-B RP signs if the subject chooses the one on the right side, regardless of how difficult is the choice for the subject. In this case, it would be reasonable to suppose that the unconscious RP determined the decision (or it was the decision itself).[11]

Nevertheless, even this untraditional Libet-style experiment has its limits. It is easy to see that it cannot refute those libertarian theories according to which it is not decisions about actions but decisions about what is good that are the source of freedom. This experiment is still a Libet-style experiment focusing on decisions about actions. The subjects previously formed their opinion on what things are desirable or good, and this is why they could give a definite answer to the question of how much they wanted to support different NPOs. Moreover, this kind of experiment is not able to refute hard-restrictivist libertarianism. Hard-restrictivists say that only one decision-type is directly free. Namely, decisions which are made during moral temptation. It is open to the hard-restrictivist to claim that these decisions are not determined by unconscious brain-states, even if all other decisions are.

For practical reasons, I do not believe that Libet-style experiments could refute hard-restrictivism (Campbell 1938/2013). The main problem is that it is needed that the subjects do not know about the real aim of the experiment. If the subjects know that their decisions between morally good and bad choices are the focus of the experiment, and there is no significant reward for making the morally problematic choice, then the subjects will not be tempted to make the morally bad choice. This is because the fact that they are perceived (and probably judged) by the researchers puts such a sociological and moral pressure on the shoulders of the subjects so that they will be not tempted to do the bad thing. Moreover, promising a high reward for making the morally bad choice would be so unprofessional or unethical that such an experiment is unlikely to be approved by an ethics committee.

11 The paper was not peer-reviewed when I accessed the article (2017–12–29). Therefore, one should not consider these results so relevant yet. But it may be worth mentioning that Maoz and his colleagues do not find such a pattern of RP that researchers have found in Buridan-type cases. If Maoz's and colleagues' results are valid, they are strong evidence that deliberative decisions are different to Buridan-type decisions from a neuroscientific point of view.

For instance, let us suppose that researchers perform a version of the Milgram-experiment. The subject is invited to a research center and the researchers say to her that they are investigating the psychological effects of punishment on the learning and memory. The subject has to give an amount of electric shock to someone in another room if she gives a wrong answer. In the original experiment, the subject has no idea that the real focus of the experiment is not the effect of punishment but whether the subject will comply with the unethical order of the researcher or not. She does not know that the other person who "receives" the shocks is, in fact, a confederate of the researchers and a skilled actor who pretends that she receives actual shocks (Milgram 1963).

It is an essential feature of the experiment that the subjects do not know that their decisions and actions are the real focus of the experiment. It is easy to see why. If they knew the real research question is whether they will obey unethical orders, they would try to do their best in order to preserve their perception of themselves as morally good people. Not to mention that, in this case, it would be important for them to present themselves as morally good persons to the experimenters. In such circumstances, most people would not be tempted to act badly even if the experimenters order them to do so. They would interpret the whole experiment as a trial in which they have the opportunity to be a moral hero.

This is even more problematic if one considers the fact that the subjects have to repeat the same type of decision many times. Even if the researchers can fool the subject once or twice about the real focus of the experiment, it would be extremely hard to do this twenty or thirty times.

I have no idea how a mixed Libet-Milgram-type experiment could be performed. Of course, I cannot rule out a priori that more creative researchers than I are able to figure out a Libet-type experiment which can in principle refute hard-restrictivism. But I think the issue that I have highlighted is challenging.[12] So challenging that one can justifiably believe that it is probable that even the best researchers cannot construe such a Libet-type experiment which could in principle refute hard-restrictivism. In sum, the main problem is that the researchers have to use some kind of device to scan the brain activity of the

12 As an anonym reviewer of the paper pointed out, hard-restrictivism is so implausible that it may not worth to spend so much time and space to discuss the difficulties of devising an experimental situation to probe it. However, I think the issue of decisions under moral temptation is relevant even if one does not care about hard-restrictivism or libertarianism at all. If someone would like to do neuroscientific experiments about decisions under moral temptations, in the light of my arguments, one should conclude that it would be wise to choose a non-Libet-style design to get some relevant result.

subject, but it seems to be impossible to use such a device which is undetectable by the subjects. And the subjects would be curious about what the device does. The researchers could lie about anything but they could not plausibly deny that the aim of the device is to collect data about the subject.

Conclusion

I argued that traditional Libet-type experiments which investigate decisions in Buridan-type situations can, in principle, refute only standard (that is, non-restrictivist) centered libertarian theories and such deliberative libertarian views which claim that decision processes about actions are the main sources of freedom. Although we have to wait for the results of untraditional Libet-type experiments, they can, in principle, refute moderate-restrictivist centered libertarian theories as well because they investigate subjects in such situations in which they have motivationally relevant reasons for more than one alternative. However, Libet-type experiments are not able to refute all libertarian theories. First, there are some deliberative libertarian theories according to which indeterministic decisions (or, if you like, judgments) about what is good may be the basis of free will. These theories cannot be refuted by Libet-style experiments because the latter focus on decisions about actions. Second, they cannot refute hard-restrictivist theories which say that agents can bring forth directly free decisions only if they are morally tempted. The problem is that if the subjects know that the researchers observe their behavior and decisions, they will not be tempted to do the morally bad action unless the reward for choosing the morally bad action is high. But it would be unprofessional or unethical to perform an experiment that gives a high reward for doing something morally bad.

Acknowledgements

This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and the Hungarian Scientific Research Fund number K109638. Moreover, I would like to thank for Alfred Mele, Andrew Cameron Sims, and an anonymous reviewer for helpful notes and advices.

References

Balaguer, M. (2012). Free will as an open scientific question. Cambridge, MA: MIT Press.

Balaguer, M. (2004). Coherent, naturalistic, and plausible formulation of libertarian free will. *Noûs, 38*, 379–406.

Campbell, C.A. (1938/2013). In Defence of free will. In C.A. Campbell: *In defence of free will: with other philosophical essays*. (pp. 35–55) Routledge.

Campbell, C.A. (1957). *On Selfhood and Godhood: The Gifford Lectures Delivered at the University of St. Andrews During Sessions 1953–54 and 1954–55*. London: Allen & Unwin.

Chisholm, R.M. (1966). Freedom and action. In K. Lehrer (ed.) *Freedom and determinism*. (pp. 11–44). New York: Random House.

Clarke, R. (2003). *Libertarian accounts of free will.* New York: Oxford University Press.

Clarke, R., & Capes, J. (2017). Incompatibilist (Nondeterministic) Theories of Free Will. E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition). URL = <https://plato.stanford.edu/archives/spr2017/entries/incompatibilism-theories/>.

Deery, O. (2015). The Fall From Eden: Why Libertarianism Isn't Justified By Experience. *Australasian Journal of Philosophy*, 93(2), 319–334.

Dennett, D.C. (1978). On giving libertarians what they say they want. In D.C. Dennett: *Brainstorms: Philosophical Essays on Mind and Psychology*, (pp. 286–299) Montgomery, Vt.: Bradford Books.

Ekstrom, L. (2003). Free will, chance, and mystery. *Philosophical Studies*, *113*(2), 153–180.

Ekstrom, L. (2000). *Free will: A philosophical study. Boulder*, CO: Westview Press.

Fischborn, M. (2017). Neuroscience and the possibility of locally determined choices: Reply to Adina Roskies and Eddy Nahmias. *Philosophical Psychology*, *30*(1–2), 198–201.

Fischborn, M. (2016). Libet-style experiments, neuroscience, and libertarian free will. *Philosophical Psychology*, *29*, 494–502.

Fried, I., Mukamel, R., & Kreiman, G. (2011). Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron*, *69*(3), 548–562.

Ginet, C. (1990). *On action*. Cambridge: Cambridge University Press.

Goetz, S.C. (1988). A noncausal theory of agency. *Philosophy and Phenomenological Research*, *49*(2), 303–316.

Griffith, M. (2007). Freedom and trying: Understanding agent causal exertions. *Acta Analytica, 22*, 16–28.

Hallett, M. (2007). Volitional control of movement: the physiology of free will. *Clinical Neurophysiology, 118*(6), 1179–1192.

Hodgson, D. (2012). *Rationality+ consciousness= free will*. Oxford: Oxford University Press.

Horgan, T. (2012). Introspection about Phenomenal Consciousness: Running the Gamut from Infallibility to Impotence. In D. Smithies & D. Stoljar (ed.), *Introspection and Consciousness*, (pp. 405–422). New York: Oxford University Press.

Horgan, T. (2011). The Phenomenology of Agency and Freedom: Lessons from Introspection and Lessons from Its Limits. *Humana.Mente* 15, 77–97.

Kane, R. (2007). Libertarianism. In J.M. Fischer, R. Kane, D. Pereboom, M. Vargas: *Four views on free will*. Oxford: Blackwell Publishing. 5–44.

Kane, R. (2000). The dual regress of free will and the role of alternative possibilities. *Noûs*, *34*(s14), 57–79.

Kane, R. (1996). *The significance of free will.* New York: Oxford University Press.

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. New York: Oxford University Press.

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

Libet, B. (1985). Unconscious cerebral initiative and the role of the conscious will in voluntary action. *The Behavioral and Brain Sciences, 8*, 519–566.

Libet, B., Gleason C.A., Wright E.W., & Pearl D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): the unconscious initiation of a freely voluntary act. *Brain* 106, 623–642.

Maoz, U., Yaffe, G., Koch, C., & Mudrik, L. (2017). Neural precursors of decisions that matter–an ERP study of deliberate and arbitrary choice. *bioRxiv*, 097626.

McCann, H. (1998). *The works of agency: On human action, will, and freedom*. Ithaca: Cornell University Press.

Mele, A.R. (2011). Free Will and Science. In R. Kane (ed.), *Oxford Handbook of Free Will, 2nd edition*. Oxford: Oxford University Press.

Mele, A.R. (2009). *Effective intentions: The power of conscious will.* New York, Oxford University Press.

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

Mele, A.R. (1995). *Autonomous Agents: From Self Control to Autonomy.* New York: Oxford University Press.

Milgram, S. (1963). Behavioral Study of obedience. *The Journal of abnormal and social psychology, 67*(4), 371–378.

Miller, J., Shepherdson, P., & Trevena, J. (2011). Effects of clock monitoring on electroencephalographic activity: Is unconscious movement initiation an artifact of the clock? *Psychological Science, 22*(1), 103–109.

Nahmias, E. (2014). Is free will an illusion? Confronting challenges from the modern mind sciences. In W. Sinnott-Armstrong (ed.), *Moral psychology, Vol. 4: Free will and moral responsibility*, (pp. 1–25). Cambridge, MA: MIT Press.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2004). The Phenomenology of Free Will. *Journal of Consciousness Studies*, 11, 162–179.

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

O'Connor, T. (2000). *Persons and causes: The metaphysics of free will.* Oxford: Oxford University Press.

O'Connor, T. (1995). Agent Causation. In T. O'Connor (ed.), *Agents, Causes, and Events: Essays on Indeterminism and Free Will,* (pp. 173–200). New York: Oxford University Press.

Pink, T. (2004). *Free will: A very short introduction* (Vol. 110). Oxford: Oxford University Press.

Reid, T. (1788). *Essays on the Active Powers of Man*. Edinburgh, Bell & Robinson.

Roskies, A.L. (2011). Why Libet's studies don't pose a threat to free will. In W. Sinnott-Armstrong (ed.), *Conscious will and responsibility*, (pp. 11–22). New York: Oxford University Press.

Roskies, A. (2006). Neuroscientific challenges to free will and responsibility. *Trends in cognitive sciences, 10*(9), 419–423.

Roskies, A., & Nahmias, E. (2017). "Local determination", even if we could find it, does not challenge free will: Commentary on Marcelo Fischborn. *Philosophical Psychology, 30*(1–2), 185–197.

Shields, G.S. (2014). Neuroscience and conscious causation: Has neuroscience shown that we cannot control our own actions?. *Review of Philosophy and Psychology, 5*(4), 565–582.

Schurger, A., Sitt, J.D., & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences, 109*(42), E2904–E2913.

Soon, C.S., Brass, M., Heinze, H.J., & Haynes, J.D. (2008). Unconscious determinants of free decisions in the human brain. *Nature neuroscience, 11*(5), 543–545.

Soon, C.S., He, A.H., Bode, S., & Haynes, J.D. (2013). Predicting free choices for abstract intentions. *Proceedings of the National Academy of Sciences, 110*(15), 6217–6222.

Steward, H. (2012). *A metaphysics for freedom*. Oxford, Oxford University Press.

Swinburne, R. (2011). Dualism and the determination of action. In R. Swinburne (ed.), *Free will and modern science,* (pp. 63–83). Oxford: Oxford University Press.

Taylor, Richard (1992). *Metaphysics* (4. ed.), Englewood-Cliffs, Prentice-Hall.

Trevena, J., & Miller, J. (2010). Brain preparation before a voluntary action: Evidence against unconscious movement initiation. *Consciousness and cognition, 19*(1), 447–456.

Van Inwagen, P. (1989). When is the will free?. *Philosophical Perspectives, 3*, 399–422.

Walter, H. (2011). Contributions of Neuroscience to the Free Will Debate: From random movement to intelligible action. In R. Kane (ed.), *Oxford Handbook of Free Will, 2nd edition*. Oxford: Oxford University Press.

Wegner, D. (2002). *The illusion of the conscious will.* Cambridge, Mass: MIT Press.

CHAPTER 6

Actions and Intentions

Sofia Bonicalzi

1 Introduction

Results in the cognitive neuroscience of volition and action have been often dismissed as ultimately irrelevant, or too week at best, to legitimately tackle the philosophical issues of free will and intentional agency. By contrast, this chapter seeks to promote a more constructive perspective regarding how philosophy and cognitive neuroscience can jointly improve our comprehension of intentional agency.

The chapter is divided into seven sections. In Section 2, I present the causal theory of action as the best attempt to provide a reductive philosophical characterisation of intentional action, introducing some early and ongoing debates concerning the causal role of conscious mental states. In Section 3, I discuss how specific problems for the understanding of intentional agency, as inherited from the causal theory, originate from widely discussed pieces of empirical evidence on how voluntary processes unfold. In Section 4, I go through some of the counter-arguments that have been put forward in order to defend the classic view of intentional agency. To a various extent, these counter-arguments target the lack of ecological validity of widely employed experimental paradigms. In Section 5, I present counter-arguments of a different type, which are based on the underlying claim that no clear causal link between unconscious neural antecedents and actions can be established on the basis of neuroscientific data.

The preoccupation expressed by some of these criticisms is shareable. Nonetheless, I will suggest that the following argument is unwarranted: *Because* it does not straightforwardly rule out the causal theory, the neuroscientific angle is irrelevant to understanding intentional agency. In Section 6, I in fact argue in favour of a different approach concerning the relation between experimental research and philosophical analysis. In particular, I suggest that the former does not simply have the role of validating the latter, but plays a more constructive part in defining the common research target. I articulate these claims with some proposals and examples (Subsections 6.1 and 6.2). Some final remarks are presented in Section 7.

- 978-90-04-40996-5

2 Ups and Downs of the *Causal Theory of Action*

The problem of naturalising intentional agency, i.e., developing an account of intentional agency that matches the scientific view of the world, remains a crucial challenge for the philosophy of action. In this respect, the *causal theory of action* (Davidson 1963; Searle 1983; Setiya 2007) has taken the lion's share, and has been considered the most suitable strategy for naturalising intentional agency in a reductive fashion. Notoriously, its central tenet is that actions are intentional if they are appropriately caused by the right mental states, such as desires-beliefs pairs (Dretske 1988) or temporally extended plans (Bratman 2000), working as the reasons for those actions. Intentional actions are thus intrinsically complex compounds, constituted by bodily movements and conscious intentional states. Indeed, the presence of intentional states distinguishes intentional actions from mere bodily movements, such as unintentional reflexes or spasms. A basic requirement in the architecture of the causal theory is that intentional states are causally relevant in virtue of their specific content (Pacherie 2011), being both relatively abstract or context-independent and linked to the overt bodily movement in a motor specific way: It is the content of my conscious abstract desire to drink water that causes the movement of my arm towards that glass on the table.

The causal theory has been held as widely appealing on at least three different levels. First, by positing an event-causal link between mental states and the corresponding actions, the causal theory does not imply the adoption of what Strawson (1962) referred to as *panicky* metaphysical views, such as agent-causal accounts (i.e., the agent, as an irreducible substance, causes events (O'Connor 2009)) or non-causal accounts (i.e., free intentional actions are uncaused or at least non-deterministically caused (Ginet 2008; Sehon 2016)). Given such a reductive aim - mental states are meant to be somehow instantiated by physical states -, the causal theory has been also seen as a good heuristic device in order to formulate psychological and computationally tractable models of intentional behaviour (Meltzoff 1995; Haynes 2007). Second, by stipulating the causal origination of intentional actions from conscious mental states, the causal theory is compatible with the seemingly transparent phenomenology of experience, whereby intentional actions are perceived as distinct from involuntary and automatic movements. Indeed, the subjective feeling of a connection between mental states and external sensory feedback, through our bodily actions, is thought to decisively contribute to the sense of agency and control over the external world (Haggard 2008).

Third, the causal connection between conscious mental states and overt bodily movements, stipulated by the causal theory, displays an immediate

relevance for ethical, normative, and legal practices. In particular, the intentional bodily actions that we are supposedly in control of, as derived from conscious intents, are recognised as specific targets of evaluative considerations, including responsibility attributions (Lagnado & Channon 2008). These three levels are tightly integrated into a unified picture, whereby intentional actions, as caused by conscious mental states, elicit the subjective sense of agency and are appropriately reached by normative evaluations. In this chapter, I mostly focus on the first of these three intertwined levels, concerning the nature of the link between conscious mental states and actions. Indeed, despite its immediate intuitiveness, the simple idea that mental states are causally involved in the production of the corresponding actions has proven to be deeply problematic.

The causal theory falls prey of many criticisms. In particular, according to a widely discussed objection (*disappearing agent*), by focusing exclusively on the causal role of mental states, the causal theory is intrinsically unable to provide a distinctive role for the agent *qua* agent as the key factor in action production. Within the causal theory, the agent is reduced to a passive bystander, unable to control the string of events that happens to take place within her mental arena (Velleman 1992). The key defence put forward by the causal theorist is exactly that positing a specific role for the agent is not needed in order to distinguish intentional actions from bodily movements. Paramount to the success of the causal theory is hence specifying the nature of the causal link between conscious mental states and actions. However, the attempt to provide an adequate account of such a causal connection has not gone unchallenged. The early and ongoing debate on deviant causal chains (Davidson 1973; Mele 1997; Schlosser 2010) highlights how conscious mental states can cause the agent's behaviour in an indirect way: By triggering a bodily movement that allows the agent to achieve the very same goal she would have wanted to achieve, but without also causing the corresponding intentional action. Typical scenarios feature agents who desire to achieve a self-posited goal but, due to the psychological state elicited by the mental state they entertain, accidentally execute a bodily movement that realises exactly the same goal. The problem is that, to properly contribute to the action, the relevant mental state must work as the reason, not just as the aetiological explanation, for the action (Anscombe 1957). To move forwards, the causal theorist has thus to clarify not only the nature of the single ingredients of the compound (mental state and bodily movement), but how these two separate elements enter into a proper relationship forming an instance of intentional agency.

Furthermore, qualifying the nature of such a relationship is crucial for compatibilist theories of free will and responsibility, often relying on the theoretical

structure of the causal theory in order to justify agential control in the absence of agent causation (McKenna 2011). In particular, according to most actual-sequence compatibilist (Haji 1998) and semi-compatibilist (Fischer & Ravizza 1998) views, the agent is acting freely and/or is responsible for her action, if she acts intentionally, i.e., if she acts on the basis of the mental states constituting her reasons for action. Crucially, actual-sequence accounts contend that, for an action to be intentional, free will is not required at the level of the selection between alternatives: I am freely and/or responsibly grasping that glass of water to the extent that I intended to do so, independently of whether I could have done something else. Acting intentionally is thus often implicitly equated to acting voluntarily, namely acting under the control of the conscious (free) will. The limit of this equation emerges in the case of actions that are voluntary under a certain description (i.e., actions that are not externally imposed, differently from actions performed under coercion or duress), without being intentional in the sense prescribed by the causal theory (i.e., they are not appropriately caused by conscious mental states). Deviant causal chains constitute a key example, but this ambiguous status extends to more ordinary cases, such as instances of negligent behaviour, habitual actions, absent-minded behaviour, and possibly episodes of weakness of will (Arpaly & Schroeder 1999). Indeed, compatibilism has traditionally had a hard time in establishing whether the agent might be held responsible for willed, but not properly intentional, actions: It is unclear whether somebody can act intentionally when some sort of high order mental state does not exercise a form of authority over the corresponding behaviour.

An anti-causalist (Frankfurt 1977) having a deep influence on the development of the causal theory (Aguilar & Buckareff 2010), Frankfurt provides a hierarchical account of the authoritative role of conscious mental states in his work on the concept of a *person*. Intrinsic to the concept of a person is the capacity to identify (with) higher order mental states (i.e., the desire to have the desire to drink water) ideally exercising control over the lower level mental states (i.e., the desire to drink water) that are ultimately responsible for the agent's behaviour. In the sequence leading to action production, the authority of conscious mental states is thus directly rooted in the agent's psychological make-up: For an agent to qualify as a person, she must be able to identify (with), or endorse, the mental states from which her actions stem. The epitome of the agent who is not acting on the basis of her own free will is the unwilling drug addict who succumbs to her (*external*) desire to take drugs, and is unable to regulate her behaviour in accordance to her higher order mental states (i.e., the desire not to desire to take drugs) (Frankfurt 1971). On a similar vein, Fischer suggests that an agent is acting freely (*guidance control*), and thus

responsibly, when she is acting in accordance to what she wants to do (Fischer 1999). In the next sections, I will label this overriding force of mental states over bodily actions, central to the causal theory, as *intentional action control*.

3 Psychological and Neuroscientific Arguments against Intentional Action Control

In the last few decades, the ability of the causal theory to support intentional agency has been challenged beyond the sphere of the philosophy of action, but once more in a way that questions the nature of the relation between conscious intentions and the corresponding bodily movements. The hallmark of the debate is the attempt to reconcile the claim that conscious intentions must play a role in causing actions with a naturalistic understanding of how cognitive processes unfold. Indeed, a number of findings in cognitive neuroscience have been read as potentially undermining intentional action control. If intentional action control is meant to be crucial for intentional agency, the conclusion would be that the same arguments ruling out intentional action control would *ipso facto* rule out intentional agency. In this chapter, I mostly side with those who contend that there is not enough empirical evidence for denying intentional action control. At the same time, I argue that the empirical investigation might have a different role to play in widening our understanding of intentional agency.

Benjamin Libet's pioneering research on the timing of conscious intentions marked a turning point in the debate about action initiation (Libet et al. 1982; Libet et al. 1983). In Libet's classic study, participants were asked to flex their right wrist or the fingers of their right hand at their own pace, executing a movement (one per trial) whenever they felt a wish or an urge to do so. The paradigm thus consisted in asking participants to spontaneously decide when to move, while the specific movement they had to perform was set from the beginning of the experiment. At the same time, participants were also required to watch a clock with a dot circling around it, and to remember the time (W) when they felt the desire or the urge to make the spontaneous movement. At the end of each trial, participants reported the time W they previously memorised. The beginning of the muscular motion associated to the onset (M) of the overt bodily movement was measured by an electromyogram (EMG), while the electrical activity during movement preparation and execution was recorded through an EEG system. The electrical readings from the scalp revealed the presence of the so-called *Readiness potential* (RP), a gradual rise in the

electrical activity in the motor cortex and in the supplementary motor area of the brain in the second or so preceding the occurrence of an intentional action (Deecke 2000). For non pre-planned actions, the RP began about 550 ms (Type II) before M, and 350 ms before W. Where the action was pre-planned, the RP (Type I) was registered as beginning around 1000 ms before M. Results comparable with those obtained by Libet have been more recently obtained with the aid of different techniques, ranging from fRMI (Soon et al. 2008; Soon et al. 2013) to implanted electrodes (Fried et al. 2011). The results showed an even earlier onset of the brain activity preceding the action, beginning up to 10 s before the subjective awareness of wanting to make a movement (Soon et al. 2008).

Prima facie at least, the aforementioned findings provide a picture of intentional agency that does not match with the causal theory. The most radical theoretical conclusion one may draw from the neuroscientific data is some version of *epiphenomenalism*. This is the theoretical thesis that seemingly causally relevant conscious processes, such as intention formation or decisions, do not play a causal role in the initiation of the corresponding action (Nahmias 2014). An epiphenomenalism of this sort is notoriously advocated by psychologist Daniel Wegner. He argues that, if the time of conscious awareness is subsequent to the onset of the unconscious neural determinants of the action, conscious mental states cannot be the cause of the corresponding bodily movements. The apparent causal relevance of conscious mental states is no more than a by-product of *post-hoc* confabulatory inferences regarding action initiation. To support the thesis, Wegner discusses a wide array of experimental data, suggesting a disconnection between the conscious experience of acting and the actual performance of a bodily movement (Wegner 2002).

Empirical evidence has indeed shown that people are less able to monitor their motor performance than the phenomenology of agency suggests (Fourneret & Jeannerod 1998). In particular, after brain stimulation of motor areas, people may report having experiences of moving where no real movement occurred, or carry out actions without perceiving any sense of agency (Desmurget et al. 2009). Furthermore, the subjective sense of agency can be altered when intentional actions are modulated by external guidance (priming), typically leading people to overestimate their own self-efficacy (Linser & Goschke 2007; Moore, Wegner, & Haggard 2009). In addition, decades of research in cognitive and social psychology have highlighted that apparent rational conscious choices are often the result of automatic, non-intelligent, processes accompanied by little cognitive elaboration (Doris 2016; Nisbett & Wilson 1977). Taken together, these pieces of evidence undermine the clear-cut

picture of intentional action as portrayed by the causal theory, by casting doubt on intentional action control. Far from being a tightly integrated, phenomenologically transparent, construct, intentional agency might thus turn out to be a black box.

4 Counter-arguments Based on the Lack of Ecological Validity of the Experimental Paradigms

With few exceptions (Nadelhoffer 2011), most philosophers have harshly criticised the validity of any theoretical conclusion drawn from neuroscientific experiments investigating intentional agency. Widely shared objections emphasise the lack of ecological validity of this kind of investigation, pointing out that the adopted experimental paradigms are unsound because ultimately unable to really tackle issues such as intentional agency and free will.

One suggested limitation depends on the type of free intentional action that experimental subjects are required to perform, which is deemed as unable to elicit feelings vaguely similar to a desire to act (Mele 2009). Such a limitation would make the corresponding results hardly generalisable to daily life scenarios in which individuals are in the position of making choices between options with appreciably different consequences. Given the absence of real reasons, of the type featuring in the causal theory, motivating them to act, participants have to be artificially *instructed* to make reportable unplanned *intentional* actions in a relatively constrained time window. In sharp contrast to this model of intentional action, Balaguer suggests that individuals are truly acting freely only when they make choices between options about which they are authentically torn, such as life-changing decisions about different job opportunities (Balaguer 2009). In contrast, the decision-making context associated to neuroscientific paradigms looks more similar to a situation where I am in front of a shelf in a supermarket and have to repeatedly choose between identical boxes of cereals - something I can do almost automatically or at least without the vigilant monitoring of conscious mental states. A related interpretative problem might arise if the type of actions experimental subjects make are so low-level that they become automatic or absent-minded: As a consequence, the data might be scarcely informative regarding the neural bases of conscious intentions. In this case, the absence of early conscious awareness might indicate that participants were not actually performing a full-fledged intentional action, but a merely voluntary action. Since many voluntary actions are performed in the absence of conscious intentions to act, some critics have plausibly argued that the absence of early conscious awareness *per se*

might be considered irrelevant for denying free will (Lavazza & De Caro 2010; Nahmias 2014).

A second recurring criticism, also regarding the type of intentional actions featuring in neuroscientific experiments on voluntary processes, concerns the exclusive focus on *immediate* intentions. Many causal theorists indeed defend a two-tier account of intentions, whereby intentions concerning the immediate present are distinguishable from long-term intentions concerning the future. Examples of this distinction can be found in Searle (prior intention vs intention-in-action (Searle 1983)), Mele (distal vs proximal intention (Mele 1992)), and Bratman (future-directed intention vs present-directed intention (Bratman 1987)). For example, according to Bratman's influential theory of agency, it is the distinctively human capacity to entertain future-directed intentions that provides the necessary connection between motivation and deliberation (Bratman 2007). In a similar vein, Shaun Gallagher has suggested that higher order planning, and not single specific motor actions, may harbour free will (Gallagher 2006).

A third criticism relies on the fact that participants are mostly required, especially in experimental settings adopting Libet's paradigm, to report the time W, namely the timing at which they consciously experience they wanted to make a movement. This moment of awareness seemingly corresponds to a second-order state or meta-state (i.e., the consciousness of the wish to make a movement), rather than to a first-order intentional state (i.e., the wish to make a movement). It is nonetheless unclear whether this meta-state must have a specific causal role in order to support intentional action control. It seems indeed plausible to assume that the awareness of being in a given motivational state occurs later than the motivational state itself (Roskies 2010; see also Dennett & Kinsbourne 1992 for an early criticism about the temporal mismatch between the subjective experience of will and the perceived position of the clock hand).

By my lights, these criticisms target real deep issues in the experimental tradition initiated by Libet. However, whereas it is plausible to argue that Libet's experiments offer an impoverished, laboratory-based, representation of daily life choices, it would be inappropriate to claim that they do not offer an example of a type of intentional action people are in the position to perform. Indeed, despite the decision between identical boxes of cereals having no appreciable consequences, our intuition is that we are performing a free intentional (i.e., willed) action in grasping box A. The mere fact that we do not care about the possible consequences of an action does not turn an intentional action into an unintentional one. In this sense, one of the reasons why agency experiments mostly feature actions bearing no relevant consequences depends

on how intentional actions are often conceptualised in cognitive neuroscience: Intentional or voluntary actions, as derived from specific motor circuits in the brain, originate from internally self-posited goals. In contrast, unintentional or involuntary actions, such as reflexes or involuntary movements, are triggered by external environmental stimuli or internal automatic states of the system (Passingham et al. 2010). Internally set goals may or may not depend on higher order reasons for acting, without this necessarily affecting the motor system. In principle, there is no clear cut, non-arbitrary, divide between high complex actions, as derived from conceptual reasoning, and simple spontaneous movements. To some extent, selecting between alternatives with no appreciable differences quintessentially exemplifies spontaneous decisions. If participants decide to move at t_1 (option 1) and not at t_2 (option 2) in a Libet's experiment, such a decision cannot be attributed to any immediately noticeable feature of the two options (external signal), but is seen as more likely to derive from purely endogenous processes.

Furthermore, in response to this type of objection, some recent experimental paradigms obtained comparable results (i.e., presence of RP for voluntary movements) while creatively bypassing the paradox of artificially prescribing people to perform intentional actions. For example, Khalighinejad et al. investigated the preparatory process (RP) prior to voluntary motor actions, by adapting a perceptual decision task in which participants were required to detect the motion of a display of dots towards the left or the right side of the screen. In each trial, participants had no clue about when the dots would have started moving coherently to the right or to the left. If participants did not wish to wait anymore, they had the option to press a *skip* button and move to the next trial. Voluntary actions were then operationalised as self-initiated skip responses while waiting for the display of dots to move coherently towards the left or the right. This way, the experimental paradigm was able to elicit the performance of intentional, or at least voluntary, actions without artificially probing participants to act (Khalighinejad et al. 2018; see also Murakami et al. 2014).

Finally, it is certainly plausible that, while walking through the aisles of the supermarket, I grasp a box of cereals remaining completely absent-minded. Such an absent-minded action can be easily intended as voluntary, if not intentional. It is indeed plausible that the lack of conscious awareness at the time of the action does not turn the action into an involuntary one. The very simple actions participants are required to perform in neuroscientific experiments seem to be of the type that might be executed completely absent-mindedly (thus without conscious awareness), remaining nonetheless voluntary. However, this assumption would require not to take into account the specific

experimental setting, whereby participants are explicitly asked to pay attention to their conscious mental states in order to report them at a later stage. Despite being of the type that can be performed absent-mindedly, the choices typically made in neuroscience experiments are thus unlikely to be made absent-mindedly. Indeed, the corresponding experimental paradigms often explicitly require people to pay attention to their conscious intentions. Using Libet's temporal judgment task in an fMRI study, Lau et al. investigated the differences in the BOLD signal by comparing a condition where experimental subjects had to report the time at which they felt the intention to move *vs* a condition where they had to report the time at which they actually executed the movement. The data showed that, when participants were explicitly required to pay attention to their intentions, as it was the case in the classic Libet's task, the blood oxygenation level-dependent (BOLD) signal in the brain areas representing intentions (pre-SMA region of the medial prefrontal cortex) was enhanced (Lau et al. 2004. See also Haynes 2007). So, whereas conscious awareness might not be required for voluntary actions in general and for the type of actions performed in neuroscientific experiment in particular, its late (i.e., following the onset of the RP) appearance in cases where subjects were explicitly asked to pay attention to their intentions cannot be so easily dismissed as irrelevant.

5 Counter-arguments Based on the Lack of Causal Clarity

The RP has been interpreted as a reliable build-up of neuronal activity consistently preceding voluntary bodily movements, and thus specifically associated to motor preparation (Kornhuber & Deecke 1990). However, the potentially causal nature of the relationship between the RP, the time W, and the subsequent overt movement is still a matter of debate. Indeed, given that a clear causal linkage between the RP and the bodily movement cannot be established, the claim that bodily movements are caused by unconscious neural antecedents, and not by conscious mental states, seems to falter. In discussing whether the RP causes W, Haggard and Eimer provided evidence that the onset of the RP varies independently of W. In particular, trials where participants show an early W were characterised by a late onset of the RP, compared to trials characterised by a late W. A better candidate for *causing* W might be the *Lateralised readiness potential* (LRP), an increase in the electrical negativity in the area contralateral to the subsequent bodily movement, and reflecting the preparation of a specific movement after the action selection is made. Coherently, the LRP began earlier for actions with an early W, compared to actions

associated to a late W. This important result suggests that the origin of W is likely to be related to a specific bodily movement rather than to a general preconscious motor preparation (Haggard and Eimer 1999).

For what concerns the relation between the RP and the overt bodily movement, an open possibility is that the RP plays a preparatory role without being sufficient for the movement to occur: In the absence of a causally relevant conscious decision to move, the subject might not execute the corresponding movement, despite the presence of the RP. In this light, Pockett and Purdy observe that the RP is neither necessary (single trials and even single experimental subjects may not show any RP) nor sufficient (changes in the electrical activity of the brain strictly resembling the RP waveforms do not necessarily originate movements) for action production (Pockett & Purdy 2011). Establishing a clear causal relation between the RP and the subsequent bodily movement implies at least the presence of the RP at the level of the single trial. By contrast, Libet's findings are based on averaging a number of trials where participants performed identical spontaneous bodily movements. In case the conscious decision to move is needed for the action to occur, conscious mental states might not have a role in action initiation, but might contribute (at least in full-fledged, non absent-minded intentional actions) to deciding whether to act or not. Libet himself, with a criticisable dualistic move, hinted at a similar solution, by arguing that, in the 150 ms preceding the onset of the movement, the subject could still decide to abort the action (Libet 1985).

By delving further into this possibility, some recent lines of research have been read as potentially reconciling the received view on intentional action with findings in the neuroscience of agency. In particular, Schurger et al. 2012 have questioned the widely accepted assumption that the RP can be truly recognised as the signal, within the motor system, of planning, preparing, and initiating a voluntary action. In contrast, they observe that changes in the neural activity preceding the bodily movement may merely reflect internal physiological noise, depending on spontaneous fluctuations and thus not specifically related to motor preparation. Within the proposed model, the timing at which the motor action occurs can be explained by means of an accumulator model: the bodily movement randomly occurs when the spontaneous electrical activity reaches a threshold, corresponding to the neural decision to move. The specific shape of the RP, normally detected before voluntary bodily movements, would be no more than an artefact due to the process of averaging the relevant epoched data time-locked to the onset of the bodily movement. Interpreted as averaged noise, the RP could hardly be seen as a signal of preconscious intentional motor preparation, playing rather a dispositional and predictive role. More speculatively, Schurger et al. suggest that the neural *decision*

to move now, corresponding to the random crossing of the sensory-motor threshold, might be close in time to the timing of the awareness of wanting to make the movement (occurring about 150 ms before the action). The temporal overlapping between the neural commitment to move (the real decision) and the subjective awareness of wanting to make a movement might suggest that the received picture of intentional agency and the neuroscientific data are not so distant as it may seem.

The neuroscience of volition has devoted a considerable amount of research to identifying the point of no return, i.e., the timing after which a subject cannot prevent herself from acting (Libet 1985; De Jong et al. 1990). In a study looking at people's ability to veto their spontaneously initiated movement, Schultze-Kraft et al. (2016) successfully detected the presence of a signal like the RP in real time, at the level of the single trial. In each trial, subjects were instructed to make intentional unpredictable bodily movements unless they received a stop signal from the computer. In the first part of the study, a Brain-Computer Interface (BCI) was trained to recognise the typical RP for each subject and predict upcoming movements. In the second part of the study, the BCI predicted bodily movements in real time (by relying on neural antecedents) and sent a stop signal to the experimental subject. The results showed that, after the onset of a RP-like signal was detected, it was still possible for the subject to avoid moving in case the stop signal was sent earlier than 200 ms before EMG. This can be taken as evidence that the RP does not necessarily lead to the occurrence of the bodily movement, which can be still cancelled by the subject until a point of no return.

The results are compatible with those by Schurger et al. 2012 to the extent that they suggest that the real decision to move (or to abort the action) does not coincide with the onset of the RP but, more or less, with the timing of conscious awareness. Clearly, the point of no return and the neural decision to move are not expected to play the same causal role. While the neural decision to move might be a necessary causal component of an intentional action, the point of no return can be just a temporal threshold after which an unconsciously initiated movement cannot be aborted. However, this temporal association between the real commitment to move (i.e., neural decision to move and point of no return) and the awareness of wanting to make a movement has been hailed as having the potential to reconcile the neuroscientific findings on voluntary processes with more traditional views on intentional agency (Schurger et al. 2016).

The power to inhibit self-initiated actions is a key feature of human action control. Nonetheless, it is unclear whether the aforementioned findings provide a picture of agency that is truly compatible with the one framed by the

causal theory. Indeed, these models do not make explicit assumptions about the origin of the neural commitment to move, supposedly following the onset of the RP and preceding the overt bodily movement. Clearly, the temporal closeness between the commitment to move and the time of conscious awareness *per se* cannot be interpreted as evidence of the fact that conscious mental states cause bodily movements: Also conscious decisions might originate from unconscious brain processes. In this light, recent experimental data suggest that the decision to inhibit or delay an action is similarly driven by antecedent and unconscious brain activity (Filevich et al. 2013).

6 A Plea for a More Constructive Role to Play for Neuroscientific Findings

Overall, most of the aforementioned discussion points originate from a similar underlying theoretical position, questioning whether neuroscientific experimental findings pose a real threat to the received view about the link between mental states and actions. Such a received view, which is well represented by the causal theory, sees conscious mental states as causing intentional actions, thus allowing agents to exercise intentional action control over their behaviour. In this respect, many argue that the neuroscientific findings that seemingly contradict the causal theory are too fragile to justify a rejection of the corresponding picture of intentional agency (Pereboom & Caruso 2018).

Whereas I also agree with this conclusion (i.e., there is not enough evidence that the causal theory is wrong), I claim that the problem with this argumentative strategy is that it implicitly assumes that specific philosophical claims (e.g. about agency or free will) can be directly (dis)proven by means of empirical investigation. According to this model, philosophical theorising is necessarily prior to empirical investigation, which has an essentially (dis)confirmatory role with respect to previously formulated philosophical claims. However, while the causal theory might work as a general framework for intentional agency (i.e., conscious mental states have a role in action production) and is clearly incompatible with other proposals (i.e., epiphenomenalism), single claims that might be inferred from the theory (i.e., how mental states should behave within the system) display little empirical translatability. As a result, when taken too literally, the causal theory may get things backwards. By contrast, my starting point is that the causal theory is meant to provide an artefactual model, and not a mechanistic explanation, of voluntary processes.

In my views, the results from cognitive neuroscience are not to be treated as a touchstone for philosophical claims, but can directly contribute to fuelling philosophical theories. The aspiration of the chapter is thus to suggest that

philosophical analysis and cognitive neuroscience can actually work together, in a mutual exchange, to ameliorate our comprehension of intentional agency. This implies that, if needed, some features of intentional agency as framed by the causal theory must be set aside. To begin with, the understanding of intentional agency offered by the causal theory has little in common with the definition of intentional or voluntary processes at play in the empirical investigation. As previously mentioned, the former insists on the appropriateness (e.g. the end-state cannot be achieved through a deviant causal chain) of the causal link between intentional mental states and bodily actions. The latter more broadly distinguishes between internally generated actions, as derived from autonomously set goals, and externally prompted movements, elicited by stimuli present in the environment (Passingham et al. 2010). This theoretical incommensurability, beginning at the basic level of operational definitions, ramifies into more high level discussions about the role of conscious mental states (Mele 2010). However, in the next two subsections, I will advance some proposals and introduce a few examples about how results in cognitive neuroscience and philosophical analysis can jointly contribute to foster understanding of intentional agency.

6.1 *Multi-level Intentions*

As mentioned earlier in the chapter, the most radical theoretical conclusion one may draw from the empirical data on voluntary processes is some form of epiphenomenalism. In response to that, many have argued that the presence of unconscious neural antecedents does not prove that the conscious decision to move has no causal role. This claim might be justified by some of the pieces of evidence discussed above, both in a negative (e.g. no evidence of a direct causal connection between the onset of the RP, W, and the bodily movement) and in a positive (e.g. possible role of the neural decision to move in association to W) fashion. The emerging idea from research in cognitive neuroscience is that conscious mental states might not play a role at the time of action initiation, but could be crucially involved in other stages of action production, such as the decision to move now (Schurger et al. 2012), or the selection between abstract action alternatives (Rowe et al. 2000). This move already implies a departure from the causal theory, to the extent that action production is not seen as a substantially unitary process, firmly guided by a conscious intention to act that remains active from action planning to action execution (i.e., from my conscious abstract desire to drink water to my hand grasping that glass on the table).

Research on the cognitive architecture of the brain has shown that the process of selecting between alternatives, planning, and executing an action involves multilayer interacting structures. At the level of action selection

between alternatives that are present in the environment, the integration of sensory information and internal representation of the state of the system is the basis for computing an action plan that is ultimately realised by muscular contraction. In the case of internally generated actions, the initial intention to act is abstract in the sense that it does not include all the details of motor execution, which are fixed by specific motor programs deriving from sub-actions (i.e., stages that are intermediate between abstract intentions and motor programs). The prefrontal cortex (PFC), and in particular the lateral prefrontal cortex (LPFC), has long been identified as the area of the brain where action generation and control take place (Duncan & Owen 2000; Miller & Cohen 2001). The proposed models of the functional organisation and cognitive architecture of the LPFC generally agree on the organisation of action control along an anterior-posterior axis, subserved by different sub-areas characterised by specific functions (Badre 2008; Bunge & Zelazo 2006; Koechlin et al. 2003; Koechlin & Summerfield 2007; Fuster 2004; Petrides 2005).

In particular, the *cascade model* proposed by Koechlin et al. 2003 and Koechlin & Summerfield (2007) describes action control as a hierarchically ordered process made possible by a cascade of top-down control from rostral to caudal LPFC and premotor regions, with anterior areas devoted to deliberative, abstract, temporally extended, action control (Grafton & Hamilton 2007; Haggard 2008; Hamilton & Grafton 2007; Kilner 2011). The model proposes that different areas of such a control network in the LPFC are responsible for executive control, defined as the capacity to select specific actions in relation to goals, thus resolving the entropy or competition between multiple action representations. Throughout this multilayer system, executive control and action coordination would nevertheless be guaranteed by the tight integration of information across the various specialised prefrontal regions. In fact, each stage in the hierarchical structure both exerts control over lower level representations and is controlled by the higher stages. These different sub-regions differentiate according to various degrees of flexibility and capacity of abstraction (i.e., their capacity to generalise across sets of representations) from the immediate action context.

Action control might thus be implemented by means of a complex hierarchical structure where different area jointly contribute to producing a given end-state. Within this framework, looking for a specific intentional state as able to play a specific causal role throughout the whole process of action production looks inevitably problematic. A more promising strategy consists in revisiting the structure of the causal theory, by accommodating this more complex view about how voluntary processes are causally integrated. For example, Pacherie has argued that goal-directed behaviour involves different levels of

action specification, subserved by three types of intentions. Distal intentions (D-intention) operate at the highest, and more abstract, level, by setting up the overarching goal of an action and the appropriate sub-goals that are necessary to reach it. Proximal intentions (P-intentions) transfer the general action plan, set up at the previous level, into the current situation of action and select the appropriate motor planning. Finally motor intentions (M-intentions), corresponding to motor representations, are in charge of setting the finest parameters and values in order to execute the action, by using external sensory information. Crucially, Pacherie suggests that the content of motor intentions may not always be accessible to consciousness (Pacherie 2008, 2015).

6.2 *Disappearing Intentions*

Acknowledging the limitations of the causal theory, some philosophers have recently moved towards solutions that bypass the aforementioned theoretical criticisms (e.g. disappearing agent, deviant causal chain), by suggesting that intentional actions have to be treated as primitive, intrinsically unified, phenomena in our psychological ontology (Ford 2011; Levy 2013, 2015; O'Brien 2010). According to Levy, the reductive program advocated by the causal theory must be abandoned since intentional actions are not analysable in terms of primitive constitutive elements (i.e., intentional mental state and bodily movement). In contrast, intentional actions are to be conceived as basic. Levy defines intentional actions as bodily movements that people can stop and continue making intentionally. One of the benefits of this account is that it can be extended to voluntary, non-intentional, actions, such as instances of negligent behaviour, habitual actions, absent-minded behaviour, and possibly episodes of weakness of the will: If you absent-mindedly step on somebody's toes, you can then step back (Levy 2013). In comparison, bodily movements such as reflexes or spasms fail to satisfy the credentials for intentional or voluntary agency because you cannot intentionally decide to stop them. Abolishing the conjunctive causal representation of intentional action further favours the unified treatment of intentional bodily and mental actions - the latter being notoriously problematic within the causal theory. In fact, in the case of mental actions, a clear separation between causes and effects is hard to establish: How can I come to pay attention to a content without somehow already attending to that content? (Proust 2001).

Such a perspective is genuinely productive to the extent that it eliminates the strong dichotomy between different types of voluntary processes. General types of human voluntary behaviour lie along a continuum that goes from simple reflexes to higher complex functions (Haggard 2014), whereby the respective tokens are characterised by varying degrees of action control.

This view resonates with a scientifically reputable view of intentional agency that characterises voluntary processes in terms of freedom from immediacy (Gold and Shadlen 2007). Furthermore, a crucial role is played by the agent's ability to decide *whether* to act or not (or, as in Levy 2013, to continue to act or not) that has been clearly identified as a key element of action control (Haggard 2008). However, to avoid the pitfall of treating empirical findings as the test-bed for theoretical models, it should be specified that such a capacity to refrain from acting or stop acting, as the key feature of intentional or voluntary actions, is not to be taken in its literal sense. Indeed, as previously mentioned, we know from neuroscience that people are not able to prevent their intentional actions from occurring after a point of no return. So, paradoxically, people are acting intentionally even when they are not (anymore) in the position of preventing themselves from acting. The proposal of treating intentional actions as primitive bypasses causality (of mental states with respect to intentional actions) *tout court*: The unit for further analysis becomes the intentional action as a whole.

A related strategy consists in diminishing the centrality of intentions as discrete entities governing the physical body, without giving up on causality as such. The operation is similar to what suggested in the previous section to the extent that a partial departure from the causal theory might be required. Indeed, we might have to accommodate a wider notion of causation with little in common with the link between intentional states and actions articulated by the causal theory. In contrast with previously discussed hierarchical models of action control, Schurgher & Uithol 2015 and Uithol et al. 2014 have pursued a strategy of this sort. They suggest that the kind of information processing occurring at the level of the PFC during action control is ultimately incompatible with the representation of intentions as context-independent, inherently causal, discrete (in terms of content and functional role) entities. The authors argue in favour of a more dynamic and context-dependent model that does not fit a literal interpretation of the causal theory: The problematic step consists once again in moving abruptly from the theoretical model framed by the causal theory to the expectation that similarly contextual-free and discrete neural realisers are present in the brain.

In support of their position, the authors bring empirical evidence speaking against the thesis that context-free, discrete, states can sit at the top of the action-hierarchy. To mention one example, research by Fuster (2001) and Fuster et al. (1982, 2000) on single cell is interpreted as showing that even the most anterior parts of the LPFC (i.e., the area usually associated to high level, abstract, deliberative control) must rely on context-sensitivity to integrate information over time. Uithol et al. 2014 observe that distinguishing the

contribution of different areas of the control network according to the different capacity for abstraction is misleading. Such a distinction is modelled on the gradient that goes from abstract intention (the intention to drink water) to concrete bodily action (grasping that glass on the table) framed by the causal theory. In contrast, the differences between areas of the LPFC is to be understood in terms of the type, source, and complexity of the information that are processed and integrated in order to produce a given end-state. In particular, more anterior areas integrate information pertaining to different types and sources (e.g. information from multiple senses), while caudal areas, which are devoted to low-level motor control, have to deal with specific information within the same type (e.g. effector-specific). Action control as a whole is thus realised through the integration of interdependent control processes in continuous relationship with contextual elements: Increasing complexity does not per se deny some form of mentally-related causality, but the kind of causality at play is substantially different from the one framed by the causal theory.

7 Conclusions

As human beings we are able to interact with the environment in various distinct ways, ranging from very simple motor actions to the implementation of long-term plans. In this chapter, I have argued that the major challenge for philosophy and cognitive neuroscience is to give reason to such a variety of instances of voluntary behaviour in a coherent manner. My overall goal was not to offer a coherent story of how intentional and voluntary agency unfold, but to provide some suggestions about possible fertile research pathways at the intersection between the philosophy and the neuroscience of action.

In philosophy, the causal theory frames intentional action control in terms of the causal authority exercised by conscious mental states over actions. Most philosophers have thus denied that findings in cognitive neuroscience can represent a real threat for the classic architecture of intentional agency as such. Whereas recognising that specific criticisms to experimental paradigms are appropriate, I have advocated a different perspective, whereby empirical findings are not necessarily to be treated as the touchstone for philosophical claims, but possibly contribute to building the theory itself. In this light, moving beyond rigid dichotomies, I suggested that some elements of the causal theory (e.g. discreteness of mental states, conscious access to all types of intentional states) might have to be abandoned in favour of a more articulated and nuanced understanding of voluntary processes.

References

Aguilar, J.H., & Buckareff, A.A., Eds. (2010). *Causing human actions. New perspectives on the causal theory of action*. MIT Press.

Anscombe, E. (1957). *Intention*. Basil Blackwell.

Arpaly, N., & Schroeder, T. (1999). Praise, blame, and the whole self. *Philosophical Studies*, 93(2):161–188.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12(5):193–200.

Balaguer, M. (2009). *Free will as an open scientific problem*. MIT Press.

Bratman, M.E. (1987). *Intention, plans, and practical reason*. Cambridge University Press.

Bratman, M.E. (2000). Reflection, planning, and temporally extended agency. *Philosophical Review*, 109(1):35–61.

Bratman, M.E. (2007). *Structures of agency: Essays*. Oxford University Press.

Bunge, S.A., & Zelazo, P.D. (2006). A brain-based account of the development of rule use in childhood. *Current Directions in Psychological Science*, 15(3):118–121.

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, 60(23): 685–700.

Davidson, D. (1973). Freedom to Act. In: D. Davidson (1980), *Essays on actions and events*. Oxford University Press, 63–81.

De Jong, R., Coles, M.G., Logan, G.D., & Gratton, G. (1990). In search of the point of no return: The control of response processes. *Journal of Experimental Psychology: Human Perception and Performance*, 16(1):164–182.

Deecke, L. (2000). The Bereitschaftspotential as an electrophysiological tool for studying the cortical organization of human voluntary action. *Supplements to Clinical Neurophysiology*, 53:199–206.

Dennett, D.C., & Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences*, 15(2):183–220.

Desmurget, M., Reilly, K.T., Richard, N., Szathmari, A., Mottolese, C., & Sirigu, A. (2009). Movement intention after parietal cortex stimulation in humans. *Science*, 324(5928):811–813.

Doris, J.M. (2016). *Talking to our selves: Reflection, ignorance, and agency*. Oxford University Press.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. MIT Press.

Duncan, J. & Owen, A.M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, 23(10):475–483.

Filevich, E., Kühn, S., & Haggard, P. (2013). There is no free won't: Antecedent brain activity predicts decisions to inhibit. *PLoS One*, 8(2):e53053.

Fischer, J.M. (1999). The value of moral responsibility. *The Proceedings of the Twentieth World Congress of Philosophy*, 1:129–140.

Fischer, J.M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Ford, A. (2011). Action and generality. In: A. Ford, J. Hornsby, F. Stoutland, Eds., *Essays on Anscombe's 'Intention'*. Harvard University Press, 76–104.

Fourneret, P., & Jeannerod, M. (1998). Limited conscious monitoring of motor performance in normal subjects. *Neuropsychologi*a, 36(11):1133–1140.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, 68(1):5–20.

Frankfurt, H. (1977). The problem of action. *American Philosophical Quarterly*, 15(2):157–162.

Fried, I., Mukamel, R., & Kreiman, G. (2011). Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron*, 69(3): 548–562.

Fuster, J.M. (2001). The prefrontal cortex-An update: Time is of the essence. *Neuron*, 30(2):319–333.

Fuster, J.M. (2004). Upper processing stages of the perception-action cycle. *Trends in Cognitive Sciences*, 8(4):143–145.

Fuster, J.M., Bauer, R.H., & Jervey, J.P. (1982). Cellular discharge in the dorsolateral prefrontal cortex of the monkey in cognitive tasks. *Experimental Neurology*, 77(3):679–694.

Fuster, J.M., Bodner, M., & Kroger, J.K. (2000). Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature*, 405(6784):347–351.

Gallagher, S. (2006). Where's the action? Epiphenomenalism and the problem of free will. In: W. Banks, S. Pockett, & S. Gallagher, Eds., *Does consciousness cause behavior? An investigation of the nature of volition*. MIT Press, 109–124.

Ginet, C. (2008). In defense of a non-causal account of reasons explanations. *The Journal of Ethics*, 12(3/4):229–237.

Gold, J.I., & Shadlen, M.N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30:535–574.

Grafton, S.T., & Hamilton, A.F. de C. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science*, 26(4):590–616.

Haggard, P. (2008). Human volition: Towards a neuroscience of will. *Nature Reviews: Neuroscience*, 9(12):934–946.

Haggard, P. (2014). Intention and agency. In: M.S. Gazzaniga, & G.R. Mangun, Eds., *The cognitive neurosciences, v*. MIT Press, 875–885.

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research*, 126(1):128–133.

Haji, I. (1998). *Moral appraisability: Puzzles, proposals, and perplexities*. Oxford University Press.

Hamilton, A.F. de C. & Grafton, S.T. (2007). The motor hierarchy: From kinematics to goals and intentions. In: P. Haggard, Y. Rossetti, M. Kawato, Eds., *Sensorimotor foundations of higher cognition attention and performance* (*Attention & performance*). Oxford University Press, 381–408.

Haynes, J.D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R.E. (2007). Reading hidden intentions in the human brain. *Current Biology*, 17(4):323–328.

Khalighinejad, N., Schurger, A., Desantis, A., Zmigrod, L., & Haggard, P. (2018). Precursor processes of human self-initiated action. *Neuroimage*, 165:35–47.

Kilner, J.M. (2011). More than one pathway to action understanding. *Trends in Cognitive Sciences*, 15(8):352–357.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6):229–235.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648):1181–1185.

Kornhuber, H.H., & Deecke, L. (1990). Readiness for movement - The Bereitschaftspotential-Story. *Current Contents Life Sciences*, 33(4):14.

Lagnado, D.A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, 108(3):754–770.

Lavazza, A., & De Caro, M. (2010). Not so fast. On some bold neuroscientific claims concerning human agency. *Neuroethics*, 3(1):23–41.

Lau, H.C., Rogers, R.D., Haggard, P., & Passingham, R.E. (2004). Attention to Intention. *Science*, 303(5661):1208–1210.

Levy, Y. (2013). Intentional action first. *The Australasian Journal of Philosophy*, 91(4):705–718.

Levy, Y. (2015). Action unified. *Philosophical Quarterly*, 66(262):65–83.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8(4):529–566.

Libet, B., Gleason, C.A., Wright, E.W., Jr., & Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (Readiness-Potential): The unconscious initiation of a freely voluntary act. *Brain*, 106(Pt 3):623–642.

Libet, B., Wright, E.W., Jr., & Gleason, C.A. (1982). Readiness-potentials preceding unrestricted "spontaneous" vs pre-planned voluntary acts. *Electroencephalography and Clinical Neurophysiology*, 54(3):322–335.

Linser, K., & Goschke, T. (2007). Unconscious modulation of the conscious experience of voluntary control. *Cognition*, 104(3):459–475.

McKenna, M. (2011). Contemporary Compatibilism: Mesh Theories and Reasons-Responsive Theories. In: R. Kane, Ed., *Oxford Handbook of Free Will*, 2nd ed. Oxford University Press, 175–198.

Mele, A.R. (1992). *Springs of action*. Oxford University Press.

Mele, A.R. (2010). *Effective intentions. The power of conscious will*. Oxford University Press.

Mele, A., Ed. (1997). *The philosophy of action*. Oxford University Press.

Meltzoff, A.N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31(5):838–850

Miller, E.K., Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202.

Moore, J.W., Wegner, D.M., & Haggard, P. (2009). Modulating the sense of agency with external cues. *Consciousness and Cognition*, 18(4):1056–1064.

Murakami, M., Vicente, M.I., Costa, G.M., & Mainen, Z.F. (2014). Neural antecedents of self-initiated actions in secondary motor cortex. *Nature Neuroscience*, 17(11):1574–1582.

Nadelhoffer, T.A. (2011). The threat of shrinking agency and free will disillusionism. In: W. Sinnott-Armstrong, L. Nadel, Eds., *Conscious will and responsibility: A tribute to Benjamin Libet*. Oxford University Press, 173–188.

Nahmias, E. (2014). Is free will an illusion? Confronting challenges from the modern mind sciences. In: W. Sinnott-Armstrong, Ed., *Moral Psychology, vol. 4: Freedom and Responsibility*. MIT Press, 1–26.

Nisbett, R.E., & Wilson, T.D. (1977). Telling More than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, 84(3):231–259.

O'Brien, L. (2010). *Self-Knowing Agents*. Oxford University Press.

O'Connor, T. (2009). Agent-causal power. In: T. Handfield, Ed., *Dispositions and causes*. Oxford University Press.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107(1):179–217.

Pacherie, E. (2011). Nonconceptual representations for action and the limits of intentional control. *Social Psychology*, 42(1):67–73.

Pacherie, E. (2015). Conscious intentions. The social creation myth. In: T. Metzinger, & J.M. Windt, Eds., *Open MIND*, 29(T). MIND Group.

Passingham, R.E., Bengtsson, S.L., & Lau, H.C. (2010). Medial frontal cortex: From self-generated action to reflection on one's own performance. *Trends in Cognitive Sciences*, 14(1):16–21.

Pereboom, D., & Caruso, G.D. (2018). Hard-incompatibilist existentialism: Neuroscience, punishment, and meaning in life. In: G.D. Caruso, & O. Flanagan, Eds., *Neuroexistentialism: Meaning, morals, and purpose in the age of neuroscience*. Oxford University Press, 193–222.

Petrides, M. (2005). The rostral-caudal axis of cognitive control within the lateral frontal cortex. In: S. Dehaene, J.-R. Duhamel, M.D. Hauser, & G. Rizzolatti, Eds., *From monkey brain to human brain. A Fyssen foundation symposium*. MIT Press, 293–314.

Pockett, S., Purdy, S. (2010). Are voluntary movements initiated preconsciously? The relationships between readiness potentials, urges and decisions. In: W. Sinnott-Armstrong, L. Nadel, Eds., *Conscious will and responsibility: a tribute to Benjamin Libet*. Oxford University Press, 34–46.

Proust, J. (2001). A Plea for Mental Acts. *Synthese*, 129(1):105–128.

Roskies, A. (2010). How does neuroscience affect our conception of volition? *Annual review of neuroscience*, 33:109–130.

Rowe, J.B., Toni, I., Josephs, O., Frackowiak, R.S., & Passingham, R.E. (2000). The prefrontal cortex: response selection or maintenance within working memory? *Science*, 288(5471):1656–1660.

Schlosser, M.E. (2010). Bending it like Beckham: Movement, control and deviant causal chains. *Analysis*, 70(2):299–303.

Schultze-Kraft, M., Birman, D., Rusconi, M., Allefeld, C., Görgen, K., Dähne, S., Blankertz, B., & Haynes, J.D. (2016). The point of no return in vetoing self-initiated movements. *Proceedings of the National Academy of Sciences USA*, 113(4):1080–1085.

Schurger, A., & Uithol, S. (2015). Nowhere and everywhere: the causal origin of voluntary action. *Review of Philosophy and Psychology*, 6(4):761–778.

Schurger, A., Mylopoulos, M., & Rosenthal, D. (2016). Neural antecedents of spontaneous voluntary movement: A new perspective. *Trends in Cognitive Sciences*, 20(2):77–79.

Schurger, A., Sitta, J.D., & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences USA*, 109(42):E2904–E2913.

Searle, J. (1983). *Intentionality*. Oxford University Press.

Sehon, S.C. (2016). *Free will and action explanation. A non-causal, compatibilist account*. Oxford University Press.

Setiya, K. (2007). *Reasons without rationalism*. Princeton University Press.

Soon, C.S., He, H.A., Bode, S., & Haynes, J.D. (2013). Predicting free choices for abstract intentions. *Proceedings of the National Academy of Sciences USA*, 110(15):6217–6222.

Soon, C.S., Brass, M., Heinze, H.J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5):543–545.

Strawson, P.F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48:1–25.

Uithol, S. Burnston, D.C., & Haselager, P. (2014). Why we may not find intentions in the brain. *Neuropsychologia*, 56:129–139.

Velleman, J.D. (1992). What happens when someone acts. *Mind, New Series*, 101(403):461–481.

Wegner, D.M. (2002). *The illusion of conscious will*. MIT Press.

PART 3

Causality and Free Will

CHAPTER 7

The Mental, the Physical and the Informational

Anna Drozdzewska

1 Introduction

One of the core experiences we share as human beings is the impact our intentions and decisions have on our life. We think that the decisions we make, from the major ones like getting married, to small ones, like taking a bottle of water from the fridge on a warm and sunny day, are what causes the actual physical overt actions. However, this shared conviction is problematic for a number of reasons, including questions like: do choices exist at all given the potentially deterministic nature of the universe; are our decisions causal in the action generation or rather the physical brain activations determine all our actions. Most of these questions are parts of a larger framework, the free will problem, which, rather than being a homogenous issue, is an umbrella term for a number of interconnected problems.

In recent decades, two most popular angles, discussed in connection to free will, although this division should not be treated as exhaustive, have emerged: (1) is free will possible given the deterministic nature of the universe, and (2) can the conclusions of neuroscientific experiments truly show that our intentions are not causal in the processes of action generation, and therefore we are not free. In this chapter I will argue that these discussions often dismiss one, important component, not only needed for the possibility of free will, but, moreover, implicitly assumed by most of the positions, namely the causal efficacy of the mental. I will argue that, if mental, as mental, does not have a causal impact on the physical, free will is in dire straits or, as Fodor famously put it:

> […]if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying…, if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world.
>
> FODOR, 1990, p. 156

This chapter will have two central, interconnected, goals: (1) to argue that the causal efficacy of the mental is one of the necessary conditions for free will,

and (2) to show that understanding causality as information transfer, might clarify the mental causation debate, while at the same time, possibly avoiding the challenge of the Causal Exclusion. To achieve these goals I will first (Section 1) show why mental causation is an important, yet often overlooked, component of the free will debate. There, I will argue that the assumption of the causal efficacy of the mental is present in both philosophical discussions, as well as neuroscientific experiments. I will then discuss the different aspects of the mental causation, from the relationship between the mind and brain to the Causal Exclusion argument I will argue that in order to preserve the possibility of free will and accommodate the data from neuroscientific experiments, non-reductive physicalism is our best bet. This, however, means that we have to find a solution to one of the most notorious problems in philosophy of mind, the Causal Exclusion argument. While some potential solutions have been proposed in the past, including the famous approach by Woodward (2003), here I will hypothesize that another approach might be better suited to capture the complexities of mental to physical interactions, namely causality as information transfer. I will argue that this way of framing causality has strong advantages over other approaches as it allows us remain in Kim's original understanding of causality(as a process), captures the wide-spread approach to the brain process in neuroscience (brain as information processing), and has the potential to better explain the connection between brain and mind.

2 The Ubiquitous Assumption

Why is mental causation so important for the existence free will? In the discussions on voluntary or intentional action, it has become a standard to explain it though evoking a reason for action (e.g. Davidson, 1980), or an intention (e.g. Mele, 2009), both of which have two components: a belief and a desire. Both are often treated as goal directed (to further distinguish them from reflexes, which are not considered free), mental rather than purely physical, and not reducible to some specific brain activation. Therefore, the usual way of thinking about a voluntary action (without thinking about determinism or indeterminism) is as a result of a causal chain, where the cause of said action is either an intention, or a reason. Alternatively, a way in which an action is identified as free is through its mental cause. Depending on the time lapse between the intention and the action, we can identify distal (connected to actions far in the future), and proximal (liked to immediate action), but regardless of this, the presence of the intention as the main feature of the causal chain which results in a voluntary action seems to be necessary. All those components are

widely accepted as mental (either in ontological or purely descriptive sense), and are considered to enter the causal chain as such. While approaching our decisions as mental also goes in line with our general intuitions about ourselves, it becomes problematic when we think about how it actually connects with our physical body, a problem that has been on the philosophical radar for centuries now. The issue is further emphasized when we include the neuroscientific findings in our considerations.

In the experimental domain, the interest in the causes of voluntary actions can be traced to the experiments by Kornhuber and Deecke (1965), but more frequently it is discussed in connection to the modified version of those conducted by Libet (Libet et al. 1982; 1985). For him an action is voluntary if it meets three conditions:

> (a) it arises endogenously, not in direct response to an external stimulus or cue; (b) there are no externally imposed restrictions or compulsions that directly or immediately control the subject's initiation and performance of the act; and (c) most important, subjects feel introspectively that they are performing the act on their own initiative and that they are free to start or not to start the act as they wish.
>
> LIBET 1985, pp. 529–530

Here an action is voluntary (and therefore free) when it is not an automatic response to a stimulus coming from the agent's immediate environment, it should be under agent's control, and he/she should recognize themselves as the agent, thereby experiencing the feeling of acting. The feeling of acting out of own initiative is a key component and can be equated with making a decision to perform an action. Therefore, similarly to philosophical arguments, Liber thinks of a free action as one that is goal oriented, and is caused by a decision of an agent. Therefore if an action that an agent would consider as arising from his own decision, was in fact caused by other factors, it would not be defined as free. To further explore the origin of fully spontaneous actions (the only ones that could have the potential to be free) Libet conducted a series of experiments. He asked the participants to perform a simple overt action: move a finger or flex a wrist while looking at a dot of light projected on a clock-like surface. The participants were then asked to remember the position of the light point at the exact moment they decided to move, while the EEG machine was recording their neural activity. The results obtained by Libet showed a consistent spike of activity in the motor cortex (where the movement would be initiated) 500 ms before the onset of the movement. Yet, on average, the participants only reported their decision to move as taking place around 200 ms

before the movement. This led Libet to conclude that the causal origins of the movement lie not in the conscious decision, as he considered necessary for voluntary actions, but rather in a non-conscious brain activation. Therefore, the brain "decides" for us when to execute the movement, and free will, if it exist at all, is limited to being a veto power, exercisable during the 200 ms between the decision and the movement, over an action already underway.

Many problems with Libet's experiments have been identified, from the doubious nature of the readiness potential, to the usage of the clock, which could possibly influende the readings. However the most basic assumptions of the experiment are rarely questioned. Libet clearly thinks that in order for an action to be free, it has to be caused by something mental (the decision) as opposed to something physical (the brain activation), and this assumption has remained the standard approach in neuroscience. The definition of the component might change- from decision to intention, but its nature, as a mental event, remains present. Therefore also here, as in philosophy, the next question that comes to mind is how.

While both philosophers (e.g. Mele) and neuroscientists (e.g. Haggard) seem to agree that a voluntary action requires a specific type of cause, namely a mental one, the relation between the mental and the physical and its importance to the argument on free will is often pushed to the side. At the same time intentions are sometimes defined as [...] *a mental state, which may be associated with particular brain states* (Pacherie and Haggard, 2010, p. 70), which settles for a vague notion of association, without further explaining how such connection could impact the causal chain leading to the voluntary action. Sometimes dualism is assumed to be necessary for free will to have a chance at existence (e.g. Haggard, 2011, Desmurget and Sirigu, 2009), but those discussions are quickly dismissed (e.g. Mele 2009, 2014). Yet, Mele himself argues that *Whenever human beings perform an overt intentional action, at least one of the following plays a causal role in its production: some intention of theirs; the acquisition or persistence of some intention of theirs; the physical correlate of one or more of the preceding items*. (Mele, 2009, p. 11), thereby assuming that it is possible for a voluntary action to be caused by the physical item, only if it is correlated with the mental one.

The main advantage of Mele's position is the space it leaves to incorporate the conclusions of neuroscientific experiments into a broader philosophical debates. It shows that the sole fact there is a physical activation linked to the mental is not in fact that surprising (after all we all have a brain), and that the physical can have a role in the causal chain that results in a free action, without automatically disproving free will. This can happen because of the relationship between the mental and the physical. The main disadvantage is the lack of

clarity on the type of relationship that connects the two, and, as always in the cases of nonreductive physicalism (which Mele's position is the classical example of), the susceptibility to the Causal Exclusion argument.

To summarize what we know so far:

(1) the voluntary action is most often defined though its causes- if it is mental, the action could be voluntary (it could also not be, there is a number of other conditions, the mental causation does not solely guarantee free will), if it is physical the action surely is not voluntary;
(2) this assumption is shared by both philosophers and neuroscientists alike, and seems to be the basic feature of what constitutes a voluntary action;
(3) when it comes to the relationship between the mental and the physical, dualism is not necessary, and the fact that the physical is a part of the causal chain is not problematic. The physical could still play a role, but only in so far it is connected to the mental.

If we accept what has been said so far here, it is clear that there is a need for an account that shows exactly how the mental and the physical are connected, and how the mental plays the causal role (in its connection to the physical). If it is not dualism, it is either reduction or nonreductive physicalism. Since the first one leaves the mental (as mental) powerless, and therefore puts free will in peril, it seems that nonreductive physicalism is the better option, where the mental is linked to the physical through some form of dependence relationship, i.e. supervenience or realization, but cannot be reduced to it. Mental is defined not ontologically, but functionally, through the role it plays in the causal chain leading to the action-outcome. Moreover, nonreductive physicalism allows for multiple realization of the mental, one of the notorious problems of the reductive approaches, according to which the same mental event can be realized by a variety of physical realizations. It also allows for seamless accommodation of the scientific results to the philosophical debate, without entailing that free will is not possible. However, its ultimate success depends on finding a satisfactory solution to the Causal Exclusion Challenge, and without it positions such as the one held by Mele, are not sufficient to explain the possibility of free will.

3 The Causal Exclusion Challenge

While nonreductive physicalism, as we have seen before, is the most promising option for us, the consistency of its premises was challenged by the causal exclusion argument, most famously described by Jaegwon Kim. The four premises are (1) principle of mental causation, (2) principle of irreducibility,

(3) principle of causal closure, and (4) no-overdetermination principle, which read:

(1) Some physical effects have mental causes (the principle of mental causation)
(2) Mental causes are distinct from physical causes (the principle of irreducibility)
(3) If a physical event has a cause at t, it has a sufficient physical cause at t (the principle of causal completeness of the physical or the causal closure of the physical domain, Kim, 2005)
(4) Events cannot have more than one sufficient cause occurring at a given time (the exclusion principle)

The first premise states that at least some mental events can have causal impact on the physical realm. The premise is very general and it allows for the mental to be either connected or disconnected from the physical realm, thereby being compatible with both nonreductive physicalism as well as dualism.

The second premise rejects the reduction of the mental to the physical realm. Also here the premise is very broad, and as such, is compatible with both dualism and nonreductive physicalism as the mental could also be independent from the physical.

The third premise necessitates that, if an event has a cause at all, it has a physical cause. The premise, by itself, does not limit the causes to one; therefore it allows for multiple causes.

The fourth premise of nonreductive physicalism limits the number of sufficient causes to only one, except for the cases of genuine overdetermination. This principle does not refer specifically to the realm of the mental, but rather states that each event has only one sufficient cause, while at the same time not claiming that such event has to be either mental or physical. However, when combined with the premise of the causal closure it entails that there can be only one physical cause.

Causal Exclusion has been one of the major philosophical puzzles, and we can identify two general groups of approaches: (1) to reject one of the premises of the causal exclusion, or (2) to redefine the notion of causation we are operating with. Often the two are combined, and one of the premises is rejected because the modified notion of causation allows for it (as in the case of interventionism).

List and Menzies (2017), who modified the first premise, to state that an agent's action is free only if it is caused by the agent's mental states, while other premises of the argument remained mostly unchanged, presented an approach particularly relevant to the free will problem. The authors explore which one of the premises could in fact be rejected, and conclude that the only

one we can potentially give up, without getting ultimately into more trouble, is the no-overdetermination principle. In other words, one effect could have more than one sufficient cause, meaning we can manipulate more than one event to obtain a change in the effect. In case of a voluntary action would mean it is caused by both the mental and the physical event, and each one of them is its sufficient cause. The authors argue that such understanding of the causal chain is possible thanks to an interventionist approach to causation (a cause is a cause though making a difference to the effect, most famously argued by Woodward (2003)).

The approach of List and Menzies is similar to arguments by causal compatibilists, who also reject the no-overdetermination argument, and claim that in the case of mental causation we are dealing with two sufficient causes. There are two strains of causal compatibilism, where:

(1) there are two, in some sense, independent causes, the mental and the physical, and the mental does not overdetermine the physical, but is the cause of the outcome, and

(2) the mental and the physical, while connected, do cause something together but their causal contributions differ, and, in some cases or under some understanding of causation, the dynamics between the mental and the outcome is more telling than the one between the physical and the outcome.

List and Menzies' account specifically resembles more the independent overdetermination solution. Then both the occurrence of the physical, as well as the occurrence of mental (in a world with free roaming souls), would bring about the same effect. Moore (2012) argues that the independent overdetermination solution will not work, and one of the reasons for it is the same as the one used in justifying the existence of the principle in the first place, namely the amount of overdetermination that would have to exist as a result. Mental causation is ubiquitous, and if mental overdetermines the physical, it would have to do so thousands of times a day, which seems to be against the odds. Secondly, accepting independent overdetermination, might lead to undermining the causal completeness of the physical. If the physical cause is sufficient, it seems no other cause is really necessary, however it seems permissible that, based just on this premise, two or more sufficient causes exist (depending on how we define sufficient). Yet, if the mental cause is sufficient, we could also conclude that the physical cause is not sufficient, which would violate the premise of causal closure directly. This, in turn, would result in the agent not being free once again.

Therefore we find ourselves in a predicament, we need causal efficacy of the mental for free will to be a possibility, but to get there we need to solve the

Causal Exclusion problem. What I want to propose in this chapter is an attempt to rethink the issue, and how we got there, and to modify how we understand causes and causation, which could leave us avoiding this major issue altogether.

4 The Neuronal Basis of Mental Causation

Most of the experiments in neuroscience investigate the causes of our actions by focusing on the final stages of motor planning and the action initiation processes. There, deliberation processes and distal intentions are not taken into account, presenting a picture that could be considered limited. Additionally, they focus on overt actions, and some argue that only overt actions can be voluntary (e.g. Haggard, 2011). As such, the discussions often do not aim to explain the specific interactions between mind and brain, nor do they go beyond the immediate causes.

A different way to think about the free will issue was proposed by Tse (2013). He expands the actions under consideration to include internal ones, and sees free will as a part of the whole decision process, starting from the moment of acquisition of a new piece of information. In other words free will, rather than in the last moment of action execution, would be intertwined in the larger decision making process, where a variety of steps can play a significant role in the outcome, thereby making the last moment of decision less important. What becomes much more relevant, however, is the causal efficacy of the mental, which is dealt with by Tse through his criterial causation.

Criterial causation relies on a set of criteria that control the neuronal firings, conditions that have to be met in order to activate the neuron. If a new information fulfills those criteria, the neuron will react, which will lead to a cascade of neuronal activation that can ultimately lead to an action. If it does not, the neuron remains inactive, and nothing happens. The multiple realization, an important feature of nonreductive approaches (which it seems Tse favors, since free will is based on the mental being causal, as mental), is accounted for since the […] *set of conditions on input* […] *can be met in multiple ways and to differing degrees*. (Tse, 2013, p. 22). At its core Tse's model of criterial causation is model of how an acquisition of a piece of new information can have a physical impact, and change the behavior of the whole system. This systemic change comes from the change of firing criteria of the neurons, based on a three stage model, where:

(1) new physical/informational criteria are set in a neuronal circuit on the basis of preceding physical/mental processing at time t1, in part via a

mechanism of rapid synaptic resetting that effectively changes the inputs to postsynaptic neuron,

(2) at time t2, inherently variable inputs arrive at the postsynaptic neuron, and

(3) at time t3 physical/informational criteria are met or not met, leading to postsynaptic neural firing or not. (Tse, 2013, p. 133).

Two things seem to be at play here- the information that changes the criteria according to which the neuron fires, and its physical realization, the vehicle of information. The physical realization of informational criteria happens on the neuronal level, where the information from the input triggers the neuron to fire if such criteria are met. Neurons transform, and communicate the information by changing individual action potential spikes into spikes that are being transmitted to other neurons, and place criteria for firing that can be met only by a specific subset of information, so only that subset can trigger the neuron to fire. Mental events can cause future mental and physical events by changing the criteria for firing used by neurons for the future inputs. The criteria that have to be met are somehow encoded in the incoming information, and as new piece of information has the power to change the current neuron-firing criteria. If those change, upon the subsequent arrival of a similar information piece, the neurons will react based on the new criteria. Tse illustrates this using the example of Orion, where prior to having any knowledge about the pattern that the stars of Orion create, no neurons will react to said pattern. However, once the criteria (i.e. the input looking like the constellation of Orion) is set, the next time we will see that pattern, a specific neuronal connection, or singular neuron will fire. Tse argues that the patterns of information (such as Orion) are multiply realizable, and the same content, while being realized in stars, beans or grains of sand, could trigger the neuronal reaction. At the same time, it is clear that both the physical realizer and the abstract informational content are taken to be causal on his account, thereby potentially necessitating the rejection of the no- overdetermination principle, which does not allow two sufficient causes (although it is not clear if, on Tse's account both would be taken to be sufficient). As such, Tse's proposed solution resembles views of causal compatibilists (e.g. Bennett, 2003; Moore, 2017 for an overview) and interventionists (e.g. Woodward, 2015).

However, a number of issues can be identified with Tse's approach, most notably what he means by causation is unclear: an experimental approach could suggest interventionism, the description of neuronal firings could rather suggest a mechanistic approach. And that lack of clarity in terms of what is understood as causes is, in my opinion, detrimental to fully showing how this account has the potential to explain the causal efficacy of the mental realm.

It also has other issues, including ascribing to neurons features usually bound to agents, like the ability to decide (e.g. Bishop, 2014). Moreover, even if the account could potentially show that the mental can in fact be causal, it can be insufficient as a stand-alone account of free will, as most of the decision making process would be unconscious, which, in turn, can be a major issue for components such as the control of own actions. At the same time, if linked with an appropriate understanding of what constitutes a causal link, the position could be seen as a first step to bring together nonreductive physicalism and the neuroscientific experiments, showing how mental, on a very low level, could potentially be causal. This however, alone, only allows to include the neuroscientific findings into the discussion more seamlessly but does not seem to solve the Causal Exclusion argument (it is still unclear how the mental could be causal along with the physical). Yet, Tse's account could be seen as a first step in seeing the mental as somehow connected to the notion of information, which in turn could give us further clues about the relationship between the mental and the physical.

5 Causality as Information Transfer

There are two main reasons to think that understanding causality as information transfer could be beneficial and applicable to the mind-brain debate: (1) the brain is an information processing machine, and (2) the way we talk about the mental to physical relationship seems similar to the relationship between the abstract content of information and its physical realization. According to Tse the transfer of information produces the change in physical world, at the same time not having to follow any of the energy conservation laws. In the pattern of Orion, it is not only the physical pattern of the stars, but the pattern combined with the abstract meaning that generates the change. Without the abstract content identified with Orion's pattern, while we have seen the same pattern repeatedly, we can hypothesize it generated no change in the neuronal connections. If we see the mental and the abstract content of information as similar (which seems plausible), we could argue that through the transfer of information, the mental can make an effect on the physical world, through the abstract information level. Therefore, if there are parallels between the mental and the abstract content of information, and the physical and the ealization of that content, understanding causality as information transfer could provide us with the most promising platform to discuss causal efficacy of the mental. However, such approach has two challenges: (1) if the mental is similar to the meaning of a piece of information, how can that be shown, and (2) how would

causation as information transfer be able to account for the causal efficacy of the meaning of said piece of information. Or, in other words, how we could develop an explanation that avoids the challenges of causal compatibilism and interventionism, while at the same time accounts for the causal efficacy of the mental.

Most of the accounts of information transfer come from mathematics (e.g. Shannon, 1948), but since they specifically get rid of the meaning of the piece of information as irrelevant, they would not serve our purpose. The most notable attempt in philosophy has been developed by John Collier's (1999), and while his approach is specifically described as physical causation, and as such might not suffice to describe the relationship between mind and brain, an extension of his theory to include the semantic part of the information could be added (which Collier himself seems to allow). Such extension could be offered by Floridi's (e.g. 2004, 2011) theory of semantic information, seen as orthogonal to the classic approaches to information. Recently other accounts (e.g. Illari, Russo, 2014) have emerged claiming that everything can be described informationally, therefore the position could be both highly applicable, and can be seen to avoid certain problems identified in connection to the standard production accounts (e.g. absences as I have come late to work because the bus did not arrive are problematic). While not yet accomplished, I want to argue that also in the case of mental to physical causation, causality understood as information transfer, could have potential to avoid certain issues.

Firstly, what is causality as information transfer. Collier himself defines it as […] *the transfer of a particular* […] *quantity of information from one state of a system to another. Physical causation is a special case in which physical information instances are transferred from one state of physical system to another*. (Collier, 1999, p. 215).

In his approach, Collier draws from Shannon's theory and treats information quantitatively rather than qualitatively (which in the case of mental causation would be a major problem). Causality is understood here as a transfer of some piece of information in its physical form, which could be seen as compatible with the type of processing Tse has in mind when he describes the neuronal firings, which ultimately lead to a behavior. Therefore, a particular piece of information would be transmitted between neurons, leading to a change in the firing criteria. However, the picture is not complete without the abstract part of the piece of information, its meaning.

While the initial work of Collier bears strong resemblance to Shannon's approach, in his later work Collier (2012) describes an information channel as a family of isomorphisms (also discussed by Dretske, 1999, and Floridi, 2010). Illari and Russo (2016) describe the covariance of isomorphisms as:

> [...] two systems a and b are couples in such a way that a's being (or type, or in state) F is correlated to b's being (of type, or in state) G, then such correlation carries for the observer of a the information that b is G. For example, the dishwasher's yellow light (a) flashing (F) is triggered by, and hence is informative about, the dishwasher (b) running out of salt (G) for an observed O, like Alice, informed about the correlation.
>
> ILLARI and RUSSO, 2016, p. 260

Importantly, the ismorphisms definition described by Illari and Russo involves the agent, who is informed by the already existing correlation between the flashing light of the dishwasher and low salt levels. This assumes that, at some point, the new information (the flashing yellow light) conveyed a new meaning (the low salt levels) to the agent, and following Tse's account, established a new firing criteria on the neurons. The addition of an agent, interpreting the information provided by the flashing light of the dishwasher, is new both to the information transfer accounts of Shannon and Collier, however, some form of coder and interpreter on both sides of the information channel could be assumed on both accounts. In both cases, Illari and Russo's and Collier's, isomorphism entails the existence of two systems, each consisting of a set of objects, and each of the objects, has a set of attributes. Isomorphism happens when attributes of one set of objects (like the flashing light) tell us something about attributes of another set of objects (like the low level of salt).

Up to this point the account can be used to explain how the agent gets informed, and acts based on the information. However, this by itself is insufficient to show that the mental is causal, to leave room for free will (even if we consider the whole, extended in time process that leads to action, from learning the new piece of information, to acting the next instance it comes into our attention again), or even to satisfactorily explain the agent's actions. Or, as neuroscientists have put it

> [...] even if a full specification of ordered synaptic potentials is the exclusive causal story, then, as functionalists of all sorts have long emphasized, reciting the specification would be a poor explanation of what happened, because there's nothing systematically special about these particular synaptic sequences that ties them to bearing reports form one occasion to the next.
>
> ROSS, SPURRETT, 2004, p. 615

Therefore the only physical causal chain is potentially insufficient, and the action of adding the salt to the dishwasher seems to need both: (1) the mental

knowledge, or a belief that the flashing light signifies the low level of salt, and (2) the physical flashing light, that both somehow contribute to the final action. In other words, understanding causality as information transfer can bring us closer to incorporating seamlessly the results of neuroscientific experiments into the discussions on action generation. Combined with Tse's criterial causation idea, causation understood as information transfer can give us an interesting account of how we acquire new information, and how that information is then relevant in the process of decision making, on a very low level. However, if we want to get to free will, we seem to, again, hit the Causal Exclusion question: was it the flashing light or the belief that it signifies the low salt level that made us go and add the salt to our dishwasher?

6 Do We Even Need Another Notion of Causation?

What constitutes a causal link and what connection is there between causality and explanation? These are two of the questions we have to face when we think about causation. Throughout the ages, causality has been defined in a variety of ways, from law-based regularities (e.g. Hume), to modern production accounts (causality as transfer, e.g. Dowe, 1992), mechanistic approaches (Craver, 2007) and interventionism (Woodward, 2003).

Currently the most broadly applied approach, also to the mental causation problem, is interventionism, where causation is understood as making a difference (the cause should make a difference to the appearance, or the probability distribution of the effect). In its current form it was presented by Woodward (2003), and has been often argued to be the effective solution to the Causal Exclusion (e.g. Shapiro and Sober 2007; Woodward 2015). However, as pointed by Baumgartner (e.g. 2009, 2012, 2013) the mental to physical relation (supervenience) does not allow for independent manipulation of competing causes and, at the same time, holding all other variables in the set fixed. If the two are necessary for establishing a causal link, as interventionists suggest, such link cannot be identified. There has been a significant amount of back and forth regarding the problem, with Woodward (2015) arguing that the non-causal dependencies that hold between two variables should exclude those from the requirement of independent manipulation. This could be interpreted to entail that establishing that something is a cause, in case of two variables connected by non-causal dependencies, both of which could potentially be a cause, relies on the choice of which variable to manipulate, as this would automatically exclude the other variable from consideration. If we want to determine if our reaching for the fridge was caused by our intention to get a cold beverage or by

the neuronal firings (on interventionism depending what we manipulate: the intention or the neuronal firings), that choice will determine which one would be considered a cause. Therefore, establishing that the mental is a cause is a metaphysical choice: whatever property we will choose to manipulate will determine our conclusions about the causal links. The account could therefore be used to argue against, rather than for, mental causation, when the physical supervening basis is chosen for manipulation (and the mental is excluded because of the non-causal dependencies connecting the two). As we have seen in the previous sections of this article, interventionism is currently most often used to answer the Causal Exclusion challenge. Its proponents often argue that, given the definition of causality, it is permissible to have two sufficient causes, therefore the fourth premise is not really a problem. But as I have argued, it is difficult to see how interventionists would avoid the issues that causal compatibilists face. Ultimately for some, including me, answering the exclusion challenge with a metaphysical choice just does not seem sufficient, but does the problem lie within the notion of causality? Or is there an issue with the way we think about causal chains and causes themselves?

Another potential approach, applicable especially in neuroscience, is the mechanistic one, most fully described by Craver in his 2007 book. The approach extends beyond biology and neuroscience, including the classic examples of machinery. Yet, as Craver argues, equating mechanistic causality with a notion of a machine is too restrictive to account for the advanced neuroscientific processes. However, in spite of the account's applicability in neuroscience, it is sensitive to the causal exclusion argument, mainly because of its multilevel explanations, which can refer to

> [...] the behaviors of organisms, the processing functions of brain systems, the representational and computational properties of brain regions, the electrophysiological properties of nerve cells, and the structures and conformation changes of molecules.
>
> CRAVER, 2007, p. 9

Moreover, as was already pointed out in the quote from Ross and Spurret, and since the mental is multiply realizable, there is nothing special about the specific neuronal firings. Moreover, because of the causal exclusion, it is difficult to show how the mental could be causal <u>via</u> mental, which is difficult on the mechanistic approach.

With the variety of different types of causalities, some (e.g. Psillos, 2009) have argued that what is described there, are not different causalities but rather different symptoms of the same thing, like in the case of a common cold.

This is a position already expressed by Anscombe (REF?) who argued that the notion of causality is far too general. Moreover, in other areas, e.g. medicine, some (e.g. Williamson and Russo, 2007) have argued that evidence of just one form is not enough, and causality can be established only when we have evidence of both mechanistic connection and evidence of correlation.

Given the problems encountered by the various theories of causality, especially when applied to the mental causation domain, and the potential need for a variety of evidence to even establish the causal chain, one might wonder if another notion of causality is what we need. So far in the article I have argued that yes, a new account is necessary, and seeing causality as information transfer can allow for a more natural way of thinking about causes and effects, especially when applied to mind- brain relationships. It seems that the notion of information transfer could very well fit both the neuroscientific jargon as well as the common intuitions about what happens in our brain. However to fully provide us with a picture of how this works, we would need another "level" of description- the meaning of information, and its causal powers in the causal chain. The account, while providing those two frameworks, does not require that two different types of causality are at work in generating one outcome, which would be unlikely. As such it is also compatible with the rejection of the no overdetermination principle, as, depending on how we define a cause, we are dealing with two, strictly linked, connected through supervenience, collaborating causes. But this might not be completely sufficient, since, as we have seen before, causal compatibilism (and such an account would potentially be an instance of it) falls short from providing a complete account of mental causation.

7 How to Move Forward

The discussions on mental causation, also those related to free will, seem to largely focus on two groups of issues: (1) what is causality, what does it mean to be a cause, and (2) how does the mental relate to the physical, and how it can be seen as a cause via being mental. However, it seems that the one thing missing is the question whether we are thinking about the causes themselves correctly. Also, here the information transfer account can help to clarify certain things. When we talk about a new piece of information being a cause, we never mean just the physical realization of the abstract meaning, nor do we ever talk about just the meaning. We tend to see information as one cause that consists of those two components: the physical vehicle and the meaning. Applying the same way of thinking to mental causation would leave us with one

mental-physical cause, which are not in competition, leaving one cause, with two necessary components. Such approach would differ from the compatibilist accounts, as it would get rid of the notion of two, dependent or independent, sufficient causes. It would also be allowed on the account of causation as information transfer as a very natural way of thinking about information. It would also avoid the Causal Exclusion argument, as again, no two competing causes would be there. However, while this picture is attractive, further work is needed to consider possible shortcomings of such an approach.

The article presented here had two main aims: (1) to show that the causally efficacious mental is necessary for the possibility of free will, and (2) that none of the current approaches to causality provides a satisfactory solution to the problem, when discussed in connection to free will. I have argued that understanding causality as information transfer can avoid the shortcomings of the currently popular approaches such as interventionism. However, to show that mental can be causal qua mental, rather than just because of the physical properties, more is needed, firstly, an information causation account that includes the semantic level of causation, secondly, possibly the rethinking of the notion of a cause. I have argued that the account of causation as information transfer is a more natural way to talk about mental causation than the current positions, but the real work begins now, in finding a seamless way to bring these theories together.

References

Baumgartner, M. (2009). Interventionist Causal Exclusion and Non-reductive Physicalism. *International Studies in the Philosophy of Science*, 23, 161–178.

Baumgartner, M. (2012). The Logical Form of Interventionism. *Philosophia* 40, 751–761.

Baumgartner, M. (2013). Rendering Interventionism and Non-reductive Physicalism Compatible. *DialecticaI* 67, 1–27 *and Us: Studies of Analytical Metaphysics: A Selection of Topics From a Methodological Perspective*, World Scientific Publishers, pp. 131–151.

Bennett, K. (2003). Why the exclusion problem seems intractable, and how, just maybe, to tract it. *Nous*, 37, 471–497.

Bishop, R.C. (2014). Review of the Neural Basis of Free Will, available at: http://ndpr.nd.edu/news/neural-basis-of-free-will-criterial-causation/.

Collier, J. (1999). Causation is the Transfer of Information. In: Sankey, H. (eds.) *Causation, Natural Laws and Explanation.* Dodrecht: Kluwer.

Collier, J. (2012). Information, causation and computation. In: Crnkovic G.D., Burgin, M. (eds.) *Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation*. World Scientific.

Craver, C.F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Clarendon Press: Oxford.

Davidson, D. (1980). Mental Events. In *Essays on Action and Events.* Oxford, Clarendon Press, 207–225.

Desmurget, M., Sirigu, A. (2009). A parietal-premotor network for movement intention and motor awareness. *Trends Cogn. Sci.* 13 (10), 411–419.

Dowe, P. (1992). An empiricist defense of the causal account of explanation. *International Studies in the Philosophy of Science*, 6 (2), 123–128.

Floridi, L. (2004). Outline of a Theory of Strongly Semantic Information. *Minds and Machines*, 14 (2), pp. 197–221.

Floridi, L. (2011). *Philosophy of Information*. Oxford University Press.

Haggard, P. (2011). Does brain science change our view of free will? In Swinbourne, R (eds.) *Free Will and Modern Science*. British Academy.

Illari, P., Russo, F. (2014). *Causality. Philosophical Theory Meets Scientific Practice.* Oxford University Press.

Illari, P. Russo, F. (2016). Causality and information. In Floridi, L. (eds.) *Routledge Handbook of the Philosophy of Information*. Routledge.

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press.

Kolmogorov, A. (1963). On Tables of Random Numbers. Reprinted in: (1998) *Theoretical Computer Science*, 207 (2): 387–395.

Kornhuber, H.H., Deecke, L. (1965). Hirnpotentialanderungen bei Willkiirbewegungen und passivenBewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. *Pfliigers Archivfur Gesamte Physiologie*, 284, 1–17. (in English: Changes in the brain potential in voluntary movements and passive movements in man: readiness potential and reafferent potential).

Lepore, E., McLaughlin, B.P. (1985). *Actions and Events: Perspectives on the Philosophy of Donald Davison*. Blackwell.

Libet, B., Wright, E.W., Gleason, C.A. (1982). Readiness-potentials preceding unrestricted 'spontaneous' *vs* pre-planned voluntary acts. *Electroencephalography and Clinical Neurophysiology*, 54, 322–335.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences*, 8, 529–566.

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

Mele, A. (2014). Free Will and Substance Dualism: The Real Scientific Threat to Free Will? In: Sinnott-Armstrong, W. (eds.) *Moral Psychology. Volume 4: Free Will and Moral Responsibility.* A Bradford Book. The MIT Press. 195–207.

Moore, D. (2012). Causal Exclusion and Dependent Overdetermination. *Erkenntnis*, 76 (3) 319–335.

Moore, D. (2017). Mental causation, compatibilism and counterfactuals. *Canadian Journal of Philosophy*, 47 (1), pp. 20–42.

Pacherie, E. (2015). Conscious Intentions – The Social Creation Myth. In Metzinger, T., Windt, J.M. (Eds). *Open MIND*: 29(T). Frankfurt am Main: MIND Group, available at.

Pacherie, E., Haggard, P. (2010). What are intentions? In: Nadel, L., Sinnott-Armstrong, W. (eds.) *Conscious Will and Responsibility. A tribute to Benjamin Libet*. Oxford University Press, 70–84.

Psillos, S. (2009). Causal Pluralism. In: Vanderbeeken, R., D'Hooghe, B. (eds.) *Worldviews, Science*, 379–423.

Ross, D., Spurrett, D. (2004). What to say to a skeptical metaphysician: A defense manula for cognitive and behavioral sceintists. *Behavioral and Brain Sciences*, 27 (5), pp. 603–627.

Russo, F., Williamson, J. (2007). Interpreting Causality in the Health Sciences. *International Studies in the Philosophy of Science*, 21 (2), pp. 157–170.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell Labs Technical Journal*, 27 (3).

Shapiro, L., Sober, E. (2007). Epiphenomenalism. The do's and don't's. In: Wolaters, G., Marchamer, P. (eds.) *Thinking about causes: From Greek philosophy to modern physics.* University of Pittsburg Press.

Shea, N. (2013). Neural mechanisms of decision-making and the personal level. In Fulford, K.W.M., Davies, M., Graham, G., Sadler, J., Stanghellini G., Thornton, T. (eds.) *Oxford Handbook of Philosophy and Psychiatry.* Oxford University Press.

Tse, P.U. (2013). *The Neural Basis of Free Will: Criterial Causation.* The MIT Press.

Woodward, J. (2003). *Making Things Happen.* Oxford University Press.

Woodward, J. (2015). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, 91 (2), 303–347.

CHAPTER 8

Free Will, Language, and the Causal Exclusion Problem

Bernard Feltz and Olivier Sartenaer

1 Introduction

The starting point of the present paper is the rather commonsensical idea that mental causation *qua* mental – that is, at the very least, some minimal degree of irreducible mental causation – is a necessary condition, though perhaps not a sufficient one, for free will (see e.g. List & Menzies 2017). In other words, it is the thought that one can only feel entitled to consider human actions as being freely willed insofar as, among other things, their ultimate origin – what ultimately caused them – is not wholly microphysical in nature (on the model of, let's say, a bunch of interrelated neurons firing). Much has been said in recent literature about how exactly to make sense of such a thought, essentially by addressing the question of how mentality could be said to be causally potent in spite of its undeniable dependence on a neural basis.[1] Here we venture into a different kind of exploration, more specifically related to the question of what irreducible mental causation can be or, more particularly, where it could come from. In this perspective, the particular question we would like to address is the following: could it be the case that what makes mental causation apparently so special or unique is that it is deeply rooted in complex forms – perhaps only to be found in human communication – of language? Put differently, could some form of irreducible linguistic causation be at the basis of the kind of mental causation that would be appropriate, among other things, for having free will?

Here is how we plan to deal with such a question. First, we briefly introduce the causal exclusion argument, essentially as an excuse to allow for mapping the space of the possible ways in which linguistic causation could in principle

1 A recurrent strategy in this respect is to embrace some form of causal pluralism by considering that mental causation is not of the same kind as physical causation. See e.g., mental causation as "fact causation" (Lowe 2008) or "criterial causation" (Tse 2013).

- 978-90-04-40996-5

be related to physical causation (Section 2). We then identify one (family) of these ways that would be appropriate to ground the kind of mental causation necessary for free will (Section 3). On the basis of this purely formal exercise, we then turn to the more ambitious objective of trying to provide plausible empirical support for the targeted (family of) schema(s) of linguistic-to-physical causation, building on recent neuroscientific work on neural plasticity (Section 4).

2 Linguistic-to-Physical Causation: Mapping the Conceptual Landscape

As of today, the causal exclusion argument certainly is one of the main challenges that any proponent of irreducible mental causation has to face. The moral of the argument is indeed essentially the following: between irreducible mental causation and a minimally monistic stance any neuroscientist should be attached to, a choice has to be made. This notoriously leads to a rather dramatic dilemma. Either one has to give up on genuine mental causation and, with it, free will, or one has to embrace some variation of Cartesian-style dualism, which prima facie doesn't fit that nicely into the current scientific picture of how minds are usually supposed to work in our world.

Besides vindicating this important moral, the causal exclusion argument has a useful pedagogical advantage. Insofar as it is structured around premises that usually act as demarcation lines between competing views on the mind-body relationship, it allows for giving an interesting overview of the conceptual landscape of the mind-body problem. Even more than that, for the argument can easily be generalized to any type of "higher-level" or "non-physical" causation, it can help to appreciate the diversity of available positions on any mode of articulation between *any* causal realms. Obviously, the ones that will be of interest here are the physical and the linguistic, from which the notations employed below naturally follow.

One possible way of framing the argument is to point out that the following set of four statements is inconsistent, L_i and P_i being different events understood as property instances occurring at times t_i:[2]

[*Dependence*]. Every L_i synchronically depends on an underlying P_i.

2 There currently exist several formulations of the argument. The present one is close in spirit to Kim's (2005, 39–45), where the dependence premise is cashed out in terms of nomological supervenience.

> [*Distinctness*]. Every L_i is distinct from any underlying P_i, in the sense that L_i has some causal potency of its own, "over and above" any P_i's causal potency. More precisely, L_i is a cause of some P_j (with $t_j > t_i$).
>
> [*Closure*]. Every P_j that has a sufficient cause has a sufficient cause P_i that is part of the same "P-level".
>
> [*Exclusion*]. No single event can have more than one sufficient cause at any time.

Although these four premises are mutually inconsistent, holding only three of them allows us to define a coherent picture of the linguistic/physical articulation. Four families of positions are thereby available:[3]

- Varieties of dualism deny [dependence], making higher-level causal events essentially independent from any lower-level basis. Such a view can come in an interactionist version, where [distinctness] involves, as mentioned above, some form of downward causation that crosses the inter-level ontological gap. Some versions are non-interactionist or parallelist, as they eschew any commitment to the clause of downward causation present in [distinctness].
- Varieties of reductionism deny [distinctness], making higher-level events either non-existent (eliminativist reductionism), causally redundant (retentive reductionism) or causally impotent (epiphenomenalism).[4]
- Strong emergentism denies [closure], to the effect that the recognition of irreducibly causal higher-level events renders (most versions of) physicalism false. In this kind of scenario, linguistic events, though dependent on the physical world, are ontologically different from – and are able to causally influence the course of – physical events.
- Weak emergentism denies [exclusion], construing L_i's causal action on P_j as essentially redundant with P_i's causal action on P_j. One way of cashing out this idea is through non-reductive physicalism built under the perspective of the "subset account" of powers, according to which L and P are different properties insofar as L's powers form a subset of P's powers.[5]

3 See Wilson (2015) for similar analysis, though more precise and complete.

4 Of course, that epiphenomenalism is here classified as a variety of reductionism is indicative of a causal realist stance that turns out to be a widespread background assumption in this kind of debate. We are well aware that epiphenomenalism can be considered as a form of antireductionism in other contexts (see e.g. Baysan, forthcoming).

5 See again Wilson (2015). Here we also adopt her terminology, keeping in mind that both the weak and strong varieties of emergence are to be considered in an ontological way. For a taxonomy that takes this into account, see Guay & Sartenaer (2016).

3 Linguistic-to-Physical Causation: The Strongly Emergentist Option(s)

Though rather coarse-grained, this brief overview of the ways in which linguistic and physical events can come to be causally related is enough for our present purpose of identifying a schema that could be conductive to free will. Although the importance and the very formulation of each of the four premises of the causal exclusion argument can be discussed – they are actually vividly debated in the literature – we suppose here that [dependence] and [distinctness] are non-negotiable ingredients of any reasonable account of free will.[6] As a result, and insofar as the package [dependence] + [distinctness] is coextensive with emergence (see e.g. Sartenaer 2016), we take it that some form of emergentism is a necessary condition for free will.

Of course, emergentism can come in many varieties, each one with its own commitments, stakes and problems. So a next step is to further identify the kind of emergentism that would be appropriate for free will. We already showed in the previous section that emergentism can come in (at least) two varieties – weak and strong – depending on whether it is consistent or not with [closure] or [exclusion]. Such a way of distinguishing between both families of emergentism hides what actually is their real demarcation criterion, which can be formulated through the following question: on the occasion of L_i emerging from P_i, are the causal powers of L_i also (possibly) exercised by P_i? Put differently, are the emerging higher-level powers also (possibly) present at the basal level?

If one takes the answer to these questions to be affirmative, then higher-level events, though existing and causally potent, don't add anything causally new to the world's ontology. They simply happen to do something that is in any case already done, so to speak. This first case, corresponding to weak emergentism, is consistent with Kim's "causal inheritance principle", according to which instances of realized properties have the same powers as the instances of their realizers (see e.g. Kim 1993, 326). Accordingly, weak emergentism can generally be seen as encapsulating the "preformationist" claim that powers of emergent entities are always already preformed in their bases (Sartenaer 2018). Because of this, this first option has to be discarded, as one cannot imagine making sense of the authorship condition of free will on the basis of the

6 In a nutshell, our motivation for this claim is rather commonsensical: an ideal account of free will should encapsulate (some degree of) irreducible non-physical causation – hence [distinctness] – in a way that is consistent with the widespread recognition that higher-level events are always somehow grounded in physical events – hence [dependence].

assumption that, whatever an agent freely does, his neurons (let's say) have already done it. In that kind of case, there would indeed be no principled reason for considering the agent as being free while denying that her neurons also be. If one is free, so must be the others, but since the others clearly aren't, then the first cannot be either. It thus seems, albeit somewhat unsurprisingly, that free will doesn't get along with causal redundancy.

This leaves us with the second option, available as soon as one is keen to answer the aforementioned demarcation question negatively. In that perspective, higher-level events have causal powers that are neither "inherited from" nor "preformed in" their physical dependence bases, so that they really bring causal novelty to the world's ontology – of course at the expense, as we've seen, of the causal closure of the physical world. This position has received many labels over the years, like, e.g., "property dualism" or "interactionist monism". We settle here for "strong emergentism", capturing the claim that higher-level events strongly emerge from lower-level physical events as soon as the first ones exert causal powers that the second ones cannot possibly exert ["strong" distinctness], while the very existence of the firsts depends on the existence of the seconds [dependence]. It is such a strong variety of emergentism that we take to constitute a necessary, though not sufficient, condition for free will, as it certainly allows for capturing the authorship condition in a robust way.[7]

This being said, two refinements can be made. First, it has now become rather standard to make use of another conceptual distinction in order to further compartmentalize the conceptual landscape of emergentism. As such, some variants of strong emergence would be classified as "synchronic", for the emergent higher-level property is supposed to be instantiated at the same moment as his lower-level, basal property. This is the canonical way of looking at emergence, which is implicitly at stake in the usual formulation of the causal exclusion argument as discussed in the previous section. Another variant of emergentism, recently come to the fore especially in philosophy of science, construes the emergence relation as essentially diachronic, the emergent being instantiated later than its basis. More than a mere detail, such distinction is of the utmost importance when it comes to metaphysical discussions, as the dependence relations at play in both cases are very different. Typically, while synchronic versions of emergence construe [dependence] in terms of

7 While focusing directly on mental causation rather than linguistic causation, other philosophers have already defended the idea that strong emergence is necessary for free will. See for instance O'Connor (2000) and Lowe (2008) in a libertarian, agent-causalist setting. For a compatibilist take on the issue, see Kistler (2010), where the emergence involved is somehow weaker (without being weak in the sense described here), insofar as it is associated with downward causation construed as the action of a constraint.

constitution, composition, realization or supervenience, diachronic varieties interpret it as being some sort of causation. Since we will later take as paradigmatic examples of linguistic downward causation cases where linguistic occurrences have been *produced* or *generated* by human beings, it is the diachronic version of (strong) emergence that will be of interest here, along the lines of, for example, O'Connor & Wong's (2005) dynamical theory of emergence (see Figure 8.1).

As a side remark, one can point out that recent variants of diachronic strong emergence eschew any commitment to the idea that the emergent should be of a higher-level than its basis, making emergence a concept that is not necessarily holistic in character (Humphreys 2016; Sartenaer 2018; Guay & Sartenaer 2018). Though it could be interesting to probe this line of thought further, it is not the place to do it here.

A second refinement is based on the plausible hypothesis that, in a putative case of linguistic, downwardly oriented causal influence, the emergent linguistic event[8] is not causally *sufficient* for bringing about the corresponding physical effect, for it is reasonable to expect that physical determinants are also at play. Accordingly, although strong emergentism tolerates that some physical effects are *wholly* determined by higher-level emergent causes, such a radical claim is not mandatory. As it appears in figure 8.1, it can also be the case that it is the conjunction of P_i and L_j that causes P_k, *provided that* – and *this* is the non-trivial matter on which strong emergentism essentially rests – the powers of L_j are not the same as the ones of P_i, to the extent that [closure] cannot possibly hold. Put differently, L_j is at least a necessary cause of P_k.

Before closing the present section, two remarks are in order, the second being incidentally the occasion of summing up our previous discussion. First, up to now, linguistic causation and higher-level linguistic events have been only conceived of as "linguistic" by way of stipulation. At this stage, we are totally noncommittal about any possible specificity of language that would make it fare in an idiosyncratic way in the debates about non-physical causation. As it has been clear, we simply exploited the high degree of generality of these debates, where the very nature of the higher-level realm under consideration is usually left unspecified.

Second and more importantly, our discussion has been purposively confined so far to the domain of "armchair" metaphysics, to the effect that the schema of linguistic-to-physical causation that we identified only describes a possible and coherent mode of articulation between both these causal levels.

8 The various levels of "linguistic event" will be detailed in section 4.2. At this stage, it can be considered that the concept of "utterance" is relevant.

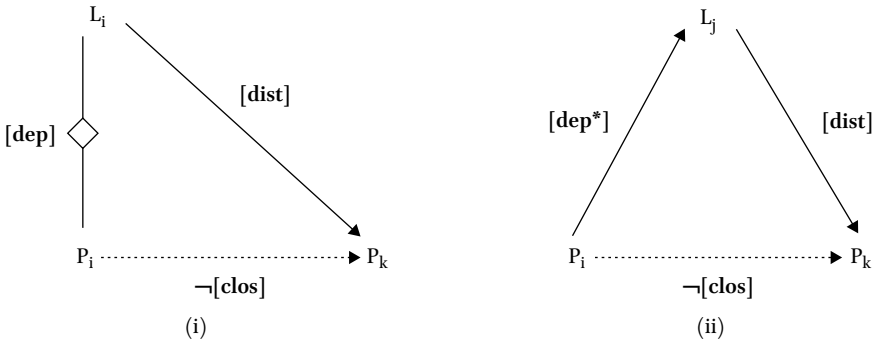

FIGURE 8.1 (i) "Traditional", synchronic strong emergence, where the conjunction of [dependence] and [distinctness] conflicts with [closure], as [exclusion] is taken to hold. (ii) Diachronic strong emergence, which only differs from (i) in that [dependence*] is construed causally. In both cases, the failure of [closure] doesn't entail that there is no causal relation at play between P_i and P_k (see main text for discussion)

As such, our endeavour has therefore been mainly descriptive and speculative. We only contended that the diachronic, strong emergence of higher-level linguistic events out of underlying physical events, together with the recognition that the associated, irreducible linguistic causation is only responsible for the advent of subsequent physical events to the extent that it acts together with underlying physical causes, is an adequate schema for making sense of the commonsensical authorship condition of free will. It is another question whether such a schema refers to any real causal dynamics in our world. This is up to science to tell us, and this is the issue to which we turn now.

4 Does Neuroscience Support the Identified Schema?

We have just seen that the schema of linguistic-to-physical causation could be a theoretical case of downward causation in a diachronic strong emergence perspective. We would like now to show how such a schema could be empirically relevant in relation with a neuroscientific point of view concerning language learning processes. Such a question has a double dimension. First of all, we would like to analyse learning processes in general, and secondly language learning processes specifically. Afterwards, in reference to philosophy of language, we would like to show how language-use in such a relation with the brain, could be the activity which permits the emergence of free will.

4.1 *Neuroscience and the Mechanisms of Learning*

Learning is a process strongly linked to memory. In all learning process, memory has an important part. Moreover, in our study of the relation between free will and causality, the relation with neuroscience is of great importance. This is why, before introducing the processes specific to language learning and downward causation, we would like first to look at the cellular mechanisms of memory.

The work of Nobel Prize winner Eric Kandel (2006) on this topic is classic. Without entering into details of the experiments, we can explain that it was in work on the mollusk Aplysia that Kandel's team first of all found evidence of two types of learning in behaviour. Certain behavioural reactions are learned for the short term – of the order of a day – while others are learned and retained for several weeks.

The researchers then focused on the mechanisms of memory at the cellular level. By studying the synaptic connections between sensory neurons and motor neurons, they found that short-term memory was linked to an increase in the release of glutamate in the synaptic cleft; this is a functional modification of the synapse. The synapse becomes more sensitive, leading to a faster stimulation of the motor neuron.

Long-term memory, on the other hand, involves an anatomical modification of the synapse, as the sensory neuron develops extensions towards the motor neuron. This process involves synthesis of proteins which requires the activity of the cell nucleus. As for short-term memory, this generation of synaptic boutons is accompanied by an increased release of glutamate, which corresponds to an increased sensitivity for the organism. Since this sensitivity takes the form of a new anatomical structure, it takes place over a much longer time-scale.

These fundamental mechanisms discovered in the context of Aplysia will be observed throughout the animal kingdom, in particular in mammals and human beings. The processes of learning vary in time. And, as in humans, short term memory and long term memory are accompanied by functional modifications and anatomical modifications, respectively.

As an example of this, Eric Kandel cites Thomas Ebert's work on musicians. With his colleagues at the University of Konstanz, Ebert's team compared images of the brains of violinists and cellists with those of non-musicians. This study revealed that the cortex area reserved for the right-hand fingers did not differ between musicians and non-musicians, whereas cortex area representations of the left-hand fingers were greatly extended in a proportion ranging from one to five in the brains of musicians. The size of a body part representation in the cortex thus depends on the intensity and complexity of its use.

Kandel draws interesting general consequences on the anthropological level: the architecture of each person's brain is unique. As soon as each behaviour has such an impact on the brain's architecture, each brain integrates the history of each individual. In consequence, even identical twins with identical genes have different brains because of their different life experiences.

Moreover, as we will see in the next section, such a mechanism can also be interpreted as a mechanism of downward causation. That concerns immediately the problematic of emergence and free will. Before this work of interpretation, we would like to refer to other mechanisms involved in learning; these mechanisms are linked to selectionist explanations.

In Neuronal Group Selection Theory (NGST), Gerald Edelman (1990, 1992, 2006), another Nobel Prize laureate, proposes a learning mechanism belonging to the logic of selectionist explanations. The genetic program induces a process of redundant connections linking sensory, motor and emotional centers within the brain. This is what Edelman calls the global cartography. This connectivity allows a multiplicity of various behaviours. An important mechanism here is selective stabilization: that implies that a nerve circuit used in the brain will be stabilized.

In this context, learning is linked to a trial and error-based strategy. Redundant structure allows for a multiplicity of behaviours. Selective stabilization mechanism in global cartography, which links sensory, motor and emotional centers, leads to the reinforcement of circuits which permit adapted behaviours. The circuits most used are reinforced to the detriment of other circuits, which remain very unstable, or even degenerate.

These second types of learning mechanism lead to the same conclusions as Kandel's mechanisms: the structure of the brain is deeply marked by personal history. The brain's fine structure is the result of the personal history of each individual.

4.2 *Learning Language: Philosophy of Language and Neuroscience*

Having introduced the neuroscientific mechanisms of learning, we would like to analyse the particular dimensions of language learning from the point of view of philosophy of language. This is in order to show how the proposed neurological mechanisms can meet these specific constraints. We will next propose an analysis which links these results to the topic of emergence and downward causation. Finally, we will demonstrate what the link between downward causation and language can contribute to our understanding of free will.

Habermas (2005) stresses at the outset that human cognition involves two dimensions: linguistic and social. Learning a language implies social

interactions. It's in this context that Habermas distinguishes "subjective mind" from "objective mind". Objective mind is "a collective knowledge preserved in symbolic form". It includes grammar, logic, semantics but also systems of meanings culturally shared. Habermas speaks about a "space of symbolically structured reasons". "The rational motivation of convictions and actions takes place in this dimension and follows logical, linguistic and pragmatic rules that are not reducible to natural laws". (Habermas, 2005, 173) (91) In this context, "subjective mind" refers to individual activity: each individual who participates in "objective mind" is on the one hand capable of understanding and using a common language, can participate in a shared conversation, can share proposed meanings and values; and on the other hand, each individual becomes able to nurture this conversation through innovation and the creation of new concepts. The learning of language is precisely the intersubjective process by which the individual appropriates "objective mind" and becomes able to implement his "subjective mind", becomes about to keep his own word and give himself his own system of meaning.

Habermas advances the hypothesis that "objective mind", the rules that structure language and the fundamental concepts of a culture, could have a structuring effect on the brain itself. In some ways, we could talk about an implementation of "objective mind" in the brain. "The meaning's systems can, in turn, influence the brains of participants through the grammatically regulated use of symbols. [...] In the course of ontogeny, the individual brain apparently acquires the dispositions required to "access" the programs of society and culture". (Habermas 2005, 175)

It's precisely this hypothesis that we would like to put forward with reference to the learning mechanisms described (among others) by Kandel and Edelman. Indeed, whether it is the functional and structural modifications or whether it is the processes of selective stabilization, learning leads to modifications of the cortex that we can interpret in Habermasian terms as an implementation of "objective mind".

This implementation allows access to "subjective mind", the individual's capability to participate in conversation which characterizes his culture. The active entry of the individual into language over the course of ontogenesis requires the learning of rules of grammar, logical rules, and the meaning of concepts. Language learning gives access to shared meanings within objective mind and makes is possible to progressively take part in the process of elaboration of meanings which characterizes all culture. In this sense, the neuroscientific mechanisms of learning constitute the empirical basis for understanding the implementation of language in an individual.

A reference to more specific linguistic studies will permit an easier relation to causation and the cellular mechanisms of learning. In a recent introduction

to *The Handbook of Language Emergence* (Wiley, Blackwell, 2015), Brian MacWhinney distinguishes "six major, partially independent hierarchies: auditory phonology, articulatory phonology, lexicon, syntax, embodied roles, and communicative structure" (MacWhinney, 2015, 4). Each of these levels concerns partially distinct neuronal areas and all these levels are interconnected. Language learning is particularly complex because learning is always linked with action. Auditory phonology is the most passive activity: a baby registers the phonemes present in his environment. That is important while it prepares the utilization of such phonemes. Articulatory phonology is already more active because it proceeds not only to repetitions of phonemes, but to articulations of phonemes conducting to the production of words. With words, we participate to the lexicon level. The integration of words in sentences implies the integration of grammar constraints, thus directly concerns syntax. A more elaborated relation to language implies an embodied role and social relations. "At the most elementary level, communicative structures involve speech acts that can then be grouped into adjacency pairs from which higher-level structures such as topic chains and narrative structure can emerge. Each of these hierarchies is tightly linked to others". (MacWhinney 2015, 4)

Habermassian implementation of "objective mind" concerns all these levels. By integrating progressively all these levels in interaction with his social environment, a child becomes able to participate to social conversations, first in a passive manner, later in an active manner. Subjective mind refers to the acquired progressive ability to have its own discourse and to participate to the social production of innovative discourses. Moreover, different meaning's systems are present in "objective mind". Each individual has to choose between a plurality of meaning's systems to organize his personal existence. These precisions in more specific linguistic terms will facilitate the articulation to the molecular mechanisms of language learning.

4.3 *Learning Language and Causation*

The moment has come to take up again the question of causation. We advance the hypothesis that language learning process mobilizes a type of downward causation in accordance with the metaphysical schema described in Section 3. The link we propose above between Habermas's conceptions of language learning and the neuroscientific mechanisms of learning indeed suggests a strong diachronic emergentism. This argument involves two dimensions: one at an ontogenetic level and one at the level of the activity of the language-user herself (Feltz 2013).

At the level of ontogenesis, things are clear: language plays a role of downward causation because these are the intersubjective interactions which lead to learning, and lead to the implementation of rules in the brain. Participation

in "objective mind" leads to learning, that is to say, the implementation of rules in the brain of the subject who becomes capable of participating in the conversation, who is then provided with "subjective mind". These processes of learning clearly reveal a dynamic which is characterized by downward causation.

More precisely, the first stages of language learning, auditory phonology, articulatory phonology, lexicon and syntax could approximately be considered as a specific form of perception. It is the case for auditory phonology. It is also partially the case for articulatory phonology and the other levels, but, in these last levels, learning is not only repeating and imitating behaviour present in environment. Learning language conducts to the ability to produce new sentences, and progressively leads an individual to be able to construct his own meaning's system. Learning language is an active process, in interaction with cultural environment, which renders able to actively participate to social discussions. That is what Habermas means by "subjective mind".

Regarding the activity of the language-user herself, things are more complex. As figure 8.2 shows, linguistic activity presupposes the learning of language (P_0); this, as we've just argued, consists in the implementation within brain connectivity of rules of grammar and meanings at time t_0. We thus presuppose that all the six levels of language learning have been realized. The brain state at time t_1 causes the sentences L_2 expressed at time t_2. Linguistic logics lead to reasoning and conceptual inventions that have an effect on the structure of the brain at P_3 so that, at time t_4, the individual expresses these new propositions L_4. In fact, we need to speak here of a continuous back and forth between language and the brain.

The image of a computer can be equally helpful here: the work of calculation takes time, the implementation of algorithms takes time, before the

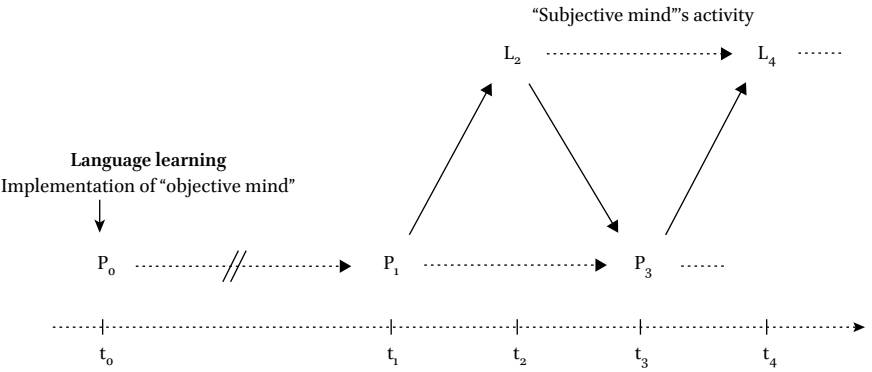

FIGURE 8.2 The diachronic, strongly emergent dynamics between the linguistic and the physical realms, as discussed in the text

production of a result that appears on the display. Linguistic activity cannot be performed independently of the activity of the brain, but the activity of language continually modifies the fine structure of the brain. "Subjective mind" is implemented by "objective mind", but permits novelty via the rules of language themselves. The brain state P_3 depends on the brain state P_1, but not entirely. It's necessary to integrate linguistic activity and its own logic in order to fully explain how state P_3 gives rise to expressions L_4.

Linguistic activity therefore takes the form of an interaction between language and the fine structure of the brain. The characteristics of language which involve the capacity for the production of innovation according to its own rules imply this continuous interaction. The impact of language via downward causation is therefore difficult to dispute in the processes of learning. To learn a language is to integrate "objective mind" into the connectivity of one's nervous system. Language has an impact on the structure of the nervous system and therefore downward causation is implied here. The place of downward causation in the use of language is more complex since we have to think of a constant interaction between language and the structure of the brain. The use of language involves the activity of the brain, but at the same time it also implies modifications in the connectivity of the brain.

Language thus has an impact on the connectivity of the brain. Language has his own logic of functioning. In a culture, language belongs to what Habermas calls "objective mind" which includes lexicon, syntax, but also "communicative structure" which includes meaning's systems. By "subjective mind", individual can participate in the evolution of "objective mind". In this sense, language can be considered as a mean to escape to a strict biological determinism. That doesn't automatically imply free will. Language could be another deterministic constraint. That is the position of many philosophers (e.g. Lacan, Levi-Strauss, Althusser or Atlan). By contrast, Habermas defends a philosophy of language where language is open to diversity in the "objective mind". Each individual has thus to make a choice between various meaning's systems present in a culture. That is "subjective mind". If language has its own logic, it can lead to escape to strict biological determinism. It can conduct to production of novelty. It can conduct each individual to adopt a specific meaning's system.[9]

9 This position could be compared with Dennett's view, which articulates free will with the activity of language and the property of linguistic reprogramming. On this point, we agree with Dennett's position. We are a little more distant concerning determinism. Dennett is compatibilist, while our approach to downward causation in relation with the neurological mechanism of learning opens the way to an uncompletely deterministic world. Addressing these issues properly would require an extensive treatment that would go far beyond the scope of this chapter.

5 Conclusion

In this paper, after having provided a general, conceptual map of the possible ways in which linguistic causation could be related to physical causation (Section 2), we proposed a metaphysically coherent picture of linguistic downward causation that would be satisfying for securing the authorship condition of free will. In particular, it was the main purpose of Section 3 to emphasize that this could be achieved under the perspective of a diachronic and strong version of emergentism, with respect to which emergent linguistic events causally depend on underlying physical events and, at the same time and in spite of such dependence, linguistic events could be the place of irreducible causal powers. We then turn in Section 4 to some recent neuroscientific works in order to assess the empirical plausibility of such a schema, supplementing the discussion with the philosophical insights of Habermas' distinction between "objective mind" and "subjective mind". In particular, it has been contended that Kandel and Edelman's works on neural plasticity offer an empirical basis for the implementation of "objective mind" and the activity of "subjective mind" under an emergentist perspective adequate for free will.[10]

References

Baysan, U. (forthcoming). Causal Emergence and Epiphenomenal Emergence. *Erkenntnis*.

Edelman, G. (1990). *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.

Edelman, G. (1992). *Bright Air, Brilliant Fire: On the Matter of the Mind*. New York: Basic Books.

Edelman, G. (2006). *Second Nature: Brain Science and Human Knowledge*. New Haven: Yale University Press.

Feltz, B., (2013). Plasticité neuronale et libre arbitre. *Revue Philosophique de Louvain*, 110, 27–52. Reedited in *Archives de philosophie du droit*, 55, Paris, 145–168.

Guay, A., & Sartenaer, O. (2016). A New Look at Emergence. Or When After is Different. *European Journal for Philosophy of Science*, 6, 297–322.

10 We would like to thank the audience of the ARC Colloquium: *Free Will, Language, and Neuroscience* held in Louvain-la-Neuve in August 2017 for helpful comments on an earlier version of this paper. Olivier Sartenaer also gratefully acknowledges the financial support of the Alexander von Humboldt Foundation.

Guay, A., & Sartenaer, O. (2018). Emergent Quasiparticles. Or How to Get a Rich Physics from a Sober Metaphysics. In O. Bueno, M. Fagan, & R.-L. Chen (Eds.), *Individuation, Process and Scientific Practices* (pp. 214–234). New York: Oxford University Press.

Habermas, J. (2005). Freedom and Determinism, in *Between Naturalism and Religion: Philosophical Essays*, 151–180. Malden: Polity Press.

Humphreys, P.W. (2016). *Emergence: A Philosophical Account*. New York: Oxford University Press.

Kandel, E. (2006). *In Search of Memory: The Emergence of a New Science of Mind.* New York: W.W. Norton & Company.

Kim, J. (1993). *Supervenience and Mind*. Cambridge: Cambridge University Press.

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

Kistler, M. (2010). Strong Emergence and Freedom: Comments on Stephan. In C. Macdonald & G. Macdonald (eds.), *Emergence in Mind*, 240–251. New York: Oxford University Press.

List, C., & Menzies, P. (2017). My Brain Made Me Do It: The Exclusion Argument Against Free Will, and What's Wrong with It. In H. Beebee, C. Hitchcock, & H. Price (eds.), *Making a Difference*, 269–285. Oxford: Oxford University Press.

Lowe, E.J. (2008). *Personal Agency: The Metaphysics of Mind and Action*. Oxford: Oxford University Press.

MacWhinney, B. (2015). *Handbook of Language Emergence*, Wiley, Blackwell.

O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.

O'Connor, T., & Wong, H.Y. (2005). The Metaphysics of Emergence. *Noûs*, 39, 658–678.

Sartenaer, O. (2016). Sixteen Years Later. Making Sense of Emergence (Again). *Journal for General Philosophy of Science*, 47, 79–103.

Sartenaer, O. (2018). Flat Emergence. *Pacific Philosophical Quarterly*, 99, 225–250.

Tse, P.U. (2013). *The Neural Basis of Free Will*. Cambridge: MIT Press.

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

Wilson, J. (2015). Metaphysical Emergence: Weak and Strong. In T. Bigaj & C. Würthrich (eds.), *Metaphysics in Contemporary Physics*, 251–306. Leiden: Brill.

Index of Authors

Index of Concepts