1 Introduction

There is a range of attitudes towards the methods of neuroethical research. Neuroethics is an interdisciplinary field that investigates the ethical significance of advancements in neuroscience for individuals and society. It is highly responsive to, and motivated by, developments in our understanding of neurology and the neurosciences.[1] Methodological views explicitly or implicitly guide neuroethical investigations. However, there are few systematic attempts at clarifying what is involved in doing neuroethics well.[2]

By focussing on exemplars of neuroethical research, our aim is to show how methodological assumptions can have a negative impact upon neuroethics.[3] Moreover, we make recommendations about how to avoid such pitfalls. By identifying these methodological problems and showing how they can be avoided we hope to chart a path for those new to neuroethics and to give those more established in this area a reason to pause for thought and reflect upon how neuroethics can thrive.

Following Adina Roskies' distinction, our discussion concerns what she calls the 'neuroscience of ethics' as opposed to the 'ethics of neuroscience'.[4] Research of the latter kind focusses upon the ethical issues raised by neuroscientific research and modifications of the brain. This type of research has sparked intense debates about the moral permissibility or desirability of moral[5] or cognitive enhancement[6] by means of actual or, more often, hypothetical interventions based on neuroscientific findings.[7] On the other hand, the neuroscience of ethics focusses on how advancements in our understanding of the neural underpinnings of behaviour might affect our views on ethical understanding and motivation and offer insight into the nature of agency.

We investigate methodological issues in the neuroscience of ethics by focussing on recent debates about responsibility. Moral and criminal responsibility and free will are perennial issues in philosophy, so prior to the enhancement debate (which enquires into the implications of using new technologies for increasing human well-being beyond remedying illness) most of the philosophical interest in neuroscience was directed towards conditions that appear to raise important questions about responsibility. Moral philosophers have also been interested in neuroscientific evidence that might lend support to rationalist or sentimentalist moral theories, so that too has been investigated in some depth.

[1] Clausen & Levy, 2015; Glannon, 2011; Illes & Federico, 2011; Levy, 2007a; Roskies, 2016.
[2] Boyle et al., 2022; Racine & Sample, 2017.
[3] In this Element, we expand upon our article 'Some methodological issues in neuroethics' (Malatesti & McMillan, 2021).
[4] Roskies, 2016. [5] Douglas, 2008; Persson & Savulescu, 2012.
[6] Savulescu & Bostrom, 2009. [7] Birks & Douglas, 2018.

We proceed as follows. In the next section, we advance the three core ideas that frame our methodological investigations. First, methodological claims concerning neuroethics should focus on how reasoned conclusions are reached in neuroethics. Second, such ways of reasoning should be described and assessed by regarding neuroethics as an interdisciplinary field and not as a discipline itself. Finally, there are at least three different aims of neuroethical reasoning, which we describe as descriptive, revisionary, and eliminative.

Having set out this general framework, we introduce an argumentative scheme (AS) that we take to characterise a central family of reasoning in neuroethics. This scheme allows us to explicate central methodological issues. Given that we apply this AS to the case study of the criminal and moral responsibility of psychopaths, we describe in detail the notion of responsibility and the construct of psychopathy. In Section 3, we focus on methodological issues and options along with the eventual costs and benefits that emerge on the ethical side of neuroethical investigations. We examine the attribution of moral and legal responsibility with a particular focus on the task of determining the psychological states and capacities that are prerequisites for them. In Section 4, we explore the conceptual issues that need attention when looking at how neuroscientific research affects our practices of holding people responsible. In Section 5, we look at the challenges related to the evidence to be used in neuroethical research in this context.

2 A Framework for the Methods of Neuroethics

2.1 Introduction

To address its methods, we need a working account of what neuroethics is. However, we do not aim to offer an exhaustive account here of what neuroethics is and its scope. The current literature offers extensive accounts of a great number of topics, main lines of debate, and results that are taken to fall within the domain of neuroethics.[8] We cannot describe and taxonomise all these debates to offer a definition of neuroethics. Instead, we will focus on a significant class of neuroethical investigations to explicate important, and often overlooked, methodological issues.

Our approach is based on the general view of neuroethics that we will set out in the following subsections. This view is based on three general assumptions. The first is that neuroethics is based on the use of arguments. The second is that because neuroethics is an interdisciplinary field its arguments involve premises that cross different disciplines. The third assumption is that these arguments cross the descriptive–normative divide.

[8] Clausen & Levy, 2015; Glannon, 2011; Illes & Federico, 2011.

These assumptions solidify in the formulation of an AS that, when applied to the case study of the legal responsibility of psychopaths, enables us to illustrate the methodological features of neuroethics.

2.2 The Centrality of Argument

Our guiding concern is how neuroethics, understood as the neuroscience of ethics, can reach reasoned convictions about the issues that fall within its scope. We thus work within a methodological tradition exemplified by the work of Henry Sidgwick, who describes the methods of ethics as 'an examination, at once expository and critical, of the different methods of obtaining reasoned convictions as to what ought to be done'.[9]

We consider the individuation and evaluation of the *types* of reasoning or arguments involved in neuroethical research. Our analysis focusses on the possible pitfalls of ways of reasoning in neuroethics.

An argument is a rational or logical process where propositions, the *premises* of the argument, are used to infer a proposition that constitutes the *conclusion* of the argument. For example, the following is a written expression of an argument:

> *Argument A.1*
> (1) If neuroscience can explain the causes of criminal behaviours, then neuroscience can be useful for the administration of criminal law.
> (2) Neuroscience can explain the causes of criminal behaviours.
> Therefore:
> (3) Neuroscience can be useful in the administration of criminal law.

The premises of the argument are the propositions (1) and (2), and proposition (3) is its conclusion. Let us now see what is involved in evaluating arguments.

An argument is said to be valid if the truth of its premises guarantees the truth of its conclusion. This means that premises offer a good support for the conclusion or, as is usually said, the conclusion follows logically from the premises. For example, *Argument A.1* is valid if, were the premises both true, the conclusion would logically follow, and the conclusion would also be true. It is worth pausing for a moment to consider in some detail what is meant by 'the validity of an argument'.

Valid arguments do not always have a true conclusion or true premises. Validity can be defined as follows:

> An argument is valid if and only if, *if* all its premises are true, *then* its conclusion must be true.

[9] Sidgwick, 1874/1981, p. vii.

This means that an argument is valid when all its premises are true and they lead us to a true conclusion. The definition of validity does not require that a valid argument necessarily has a true conclusion. It would have a true conclusion only if all its premises were true. Similarly, a valid argument does not need to have all true premises. Consider the following example of a valid argument with a false conclusion and a false premise:

Argument A.2
(1) If current neuroimaging techniques permit us to read minds, then language use is obsolete.
(2) Current neuroimaging techniques permit us to read minds.
Therefore:
(3) Language use is obsolete.

Adopting the conclusion of the argument would be a bad idea, as it is false. But the argument is valid. Were premises (1) and (2) true, the conclusion (3) would also be true. However, premise (2) is surely false. Whether premise (1) is true or false is open to further analysis and evidence, so it is less obviously false and up for debate. Let us now consider further, in more detail, what is meant by the 'validity of an argument'.

An argument's validity depends on its *logical form*. To understand the logical form of an argument, observe that the premises and conclusion of *Argument A.2* can be reformulated by means of letters, which replace the propositions:

(1) If P (current neuroimaging techniques permit us to read minds), then Q (language use is obsolete).
(2) P (current neuroimaging techniques permits us to read minds).
Therefore:
(3) Q (language use is obsolete).

Thus, we realise that the form of the argument is the following:

(1) If P, then Q.
(2) P.
Therefore:
(3) Q.

The above schematic structure exhibits the logical form of the argument. The logical form of an argument depends on the logical structure of the premises and their conclusion. For example, in the argument above, no matter the specific propositions that we can substitute instead of the variables P or Q, the first premise must have the following logical form:

(1) If P, then Q.

This logical form characterises a *conditional*, where P is called the *antecedent* and Q the *consequent*. Moreover, the second premise of the argument, must be identical to the antecedent of the first premise (1): if P then Q, that is, it must be the proposition P.

This schematic structure is *formal* because it describes abstract features that can be shared by different arguments. Consider again *Argument A.1*:

(1) If neuroscience can explain the causes of criminal behaviour, then neuroscience can be useful for criminal law.
(2) Neuroscience can explain the causes of criminal behaviour.
Therefore:
(3) Neuroscience can be useful for criminal law.

By replacing the propositions that appear in it, we obtain:

(1) If P (neuroscience can explain the causes of criminal behaviour), then Q (neuroscience can be useful for criminal law).
(2) P (neuroscience can explain the causes of criminal behaviour).
Therefore:
(3) Q (neuroscience can be useful for criminal law).

This means that this argument has the logical form:

(1) If P, then Q.
(2) P.
Therefore:
(3) Q.

Let us now look at a fundamental relation between the logical form of an argument and its validity. It is a fundamental contribution of logic that arguments are valid in virtue of their logical form. What makes *Argument A.1* and *Argument A.2* valid is the fact that they are both instances of the scheme we have introduced:

(1) If P, then Q.
(2) P.
Therefore:
(3) Q.

Formal arguments with that structure are known as *modus ponens*. Besides *modus ponens*, logicians have identified the logical form of other valid arguments. *Modus tollens* is perhaps the second most common argument. It is another fundamental, valid AS and has the following form:

(1) If P, then Q.
(2) Not Q.

Therefore:
(3) Not P.

Let us consider the following instance of *modus tollens*:

(1) If the moral capacities of individuals are enhanced with biomedical interventions upon their brain (P), then biomedical interventions upon the brain will increase their freedom of choice (Q).
(2) Biomedical interventions in the brain do not increase the freedom of individuals (not Q).
Therefore:
(3) The moral capacities of individuals are not enhanced by biomedical interventions upon their brain (not P).

There are many other valid argumentative forms, but a surprising number of arguments in neuroethics, both sound and weak, draw upon variations of *modus ponens* and *tollens*.

Our first methodological recommendation – that we should look to the methods employed in neuroethics for reaching 'reasoned convictions' – is premised upon the assumption that the arguments in this field should be valid. However, it is not enough to reflect on the definition of the validity of an argument to reach a reasoned conviction. Validity is best thought of as only a necessary condition of good neuroethics. This is partly because valid arguments generate true conclusions only when their premises are true. Thus, a more complete requirement is that good neuroethics should offer valid arguments with true premises.

An argument is said to be *sound* when it is valid, and all its premises are true. Consider the following instance of *modus ponens*:

(1) If a town is close to Lake Taupo, the town is in New Zealand.
(2) Tūrangi is close by Lake Taupo.
Therefore:
(3) Tūrangi is in New Zealand.

Being an instance of *modus ponens*, the argument is valid. Thus, if its premises are true, its conclusion must also be true. Both the premises of this argument, as a matter of geography, are true. Therefore, the conclusion is true.

We can, thus, say that an aim of good neuroethics is to offer sound arguments. This means that besides using logically valid arguments neuroethicists should attempt to demonstrate that the premises of their arguments are true. So, a central task of methodological investigation is to investigate the sources of evidence that can be used in neuroethics to support the truth of the premises used when advancing arguments. As we will show, it is the use of evidence that

is one of the principal pitfalls of neuroethics. However, before addressing this task, we must complicate further our picture of reasoning. What we have said so far concerns *deductive* arguments, that is arguments that when valid, lead from true premises to a true conclusion. But there are other important arguments that are not deductive.

The relationship between the premise and conclusion of an argument is not always deductive. For instance, consider the following premise and inference:

(1) There are paw prints in the snow.
Therefore:
(2) My cat walked in the snow.

If we own a cat, know that it might have been out and able to walk in the snow, know that there are not many other cats around that might have walked there, we might form the belief that our cat walked there. We might not attach a great deal of confidence to that belief, and view it as something that could easily be rebutted by something like the presence of other cats. Nonetheless, that kind of inference is part of how we make inferences and form beliefs every day.

This type of reasoning is an 'inductive' argument, which often involves going beyond the premises. Inductive arguments are not formally valid but can still be considered good or bad based on how reasonable the inference is. In this example, if I know there are other cats around or that my cat is old and arthritic and usually avoids walking in the snow, my inference might be bad, or poor given that I had reasons to rebut that inference.

While a full explanation of the fundamentals of logic is beyond the scope of this Element, these brief remarks are sufficient for our purpose: our emphasis upon the kind of reasoning that typifies neuroethics and what this shows about its methods.[10] We will now focus upon how evidence is gathered for the premises in the deductive and inductive arguments advanced within neuroethics.

2.3 Neuroethics as an Interdisciplinary Field

Examining the reasoning or arguments of neuroethics involves *descriptive* and *evaluative* components. The descriptive component explicates the aims that guide some types of argument, the kind of evidence that is adopted in these arguments, and the methods used to obtain it. The evaluative component of our methodological study highlights common and possible pitfalls that can afflict this kind of reasoning and argument. We will investigate how these problems derive from the inappropriate aims that are pursued in some arguments, and from defective methods used for gathering evidence. However, to appreciate the

[10] For an introduction to formal and informal logic, see Cohen et al., 2019.