Conceptual and Theoretical Issues in Forensic Neuropsychology

By

Lucy Elizabeth Jackson

A thesis

submitted to the Victoria University of Wellington

in fulfilment of the requirements for the degree of

Master of Science in Forensic Psychology

Victoria University of Wellington

2018

Abstract

Neuroscience is an increasingly popular area of study in forensic psychology, and there is a large body of empirical research emerging investigating the biological basis of offending behaviour. However, the theoretical and conceptual underpinnings of forensic neuroscience are currently underdeveloped. The aim of this thesis is to provide insight into the potential issues with forensic neuroscience and provide a number of suggestions for researchers to follow. This thesis begins by outlining these theoretical, conceptual, and empirical issues that researchers should be considering, including conceptualisation of the mind, explanation, and the methodological issues in neuroscience. These issues are then examined in more detail using two specific subject areas as exemplars: deception detection and mind-reading by brain-reading. This thesis concludes with suggestions for future researchers, which include making sure that research is based on a strong theoretical framework, clarity around the kind of explanation employed and use of explanatory pluralism, clear and consistent definitions to improve conceptual validity, using consistent and conceptually valid experimental protocols, and explicit consideration of technical limitations and how they impact the validity of the experiment.

Acknowledgements

I would first like to thank Professor Tony Ward for supervising this thesis. Your extensive knowledge, great enthusiasm for all topics, and support have been invaluable in the creation of this thesis.

Thank you to all the staff in the forensic psychology program for giving me the opportunity and guidance to discover my passion.

Thank you to all the students in both the Forensic Lab and the EPC Lab for all of your support in helping me reach my goals, and all of the inspirational talks and conversations! A special thanks to Roxy for her help in editing this thesis.

Thank you to all my friends here at Victoria University. While it can be intimidating to uproot your life and move to a new city, you have all showed me such great support and amazing friendship throughout the past two years.

Finally, thank you to my family for their unconditional love and support. Particularly thank you to my parents. You have continued to support me in following my passions, no matter where they take me. I am eternally grateful for the love and guidance that you show me.

Table of Contents

List of Figures

List of Tables

CHAPTER 1: Introduction

While criminal behaviour has long been of interest within psychology, the sub-field of forensic psychology is still a relatively new and evolving area. The Psychology of Criminal Conduct (PCC) remains the most influential model for understanding offender psychology and its relationship to crime (Andrews & Bonta, 2010). Having originally been developed from sociological and criminological theories such as control theory, social learning theory, and strain theory, it continues to lean heavily on social factors involved in offending behaviour (Andrews & Bonta, 2010). This is reflected in the risk, need, responsivity (RNR) model that guides correctional policy, assessment and treatment in the majority of offender treatment programs around the world. This model states that treatment targets should be *dynamic* (changeable) *risk factors*: aspects of a person or a person's life that increase the risk of offending. These risk factors are largely centred on the social, behavioural, and cognitive levels of explanation. Examples of these factors include antisocial associates, substance use, antisocial cognitions, and school or work circumstances (Andrews & Bonta, 2010).

Though these risk factors certainly provide a useful metric for determining risk for offenders and identifying potential treatment targets, in recent years there has been an increasing interest in integrating neuroscientific research into the field. Neuroscience is an area of study that focuses on the anatomy and function of the brain and central nervous system. Cognitive neuroscience is more specifically interested in the relationship between the mind and the brain; in how the brain can (seemingly) create the mind and control our behaviour (Uttal, 2011). Experiments in this area focus on the cellular/molecular levels of analysis, patterns of activity throughout the brain, and anatomical differences between subjects. Cognitive neuroscience has rapidly increased in popularity over the past few decades due to continual advances in neuroscientific technology and techniques for analysing the brain. These include functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) as the most common techniques, but other techniques such as positron emission tomography (PET) and, more recently, transcranial magnetic stimulation (TMS) are also in use. The majority of neuroscientific experiments use functional magnetic resonance imaging (fMRI) techniques. Here, the blood-oxygen level dependant (BOLD) signal is detected by measuring the changes in the blood flow in the brain. Statistical analysis, including voxel or pattern analysis, can then be carried out to determine which areas are significantly differentially activated, when compared to the subject's brain at rest. Researchers can then transpose this activity onto a canonical map of the brain, to show activity levels in certain areas.

In forensic psychology the interest in cognitive neuroscience has increased considerably in recent years (Ward, Wilshire & Jackson, 2018). This includes a special issue on neuroscience recently published by *Psychology, Crime and Law*, and a comprehensive 36-chapter handbook on a wide range of topics related to forensic neuroscience (Beech et al., 2018). Research has focused on discovering the neurobiological correlates of offending behaviour, in order to produce more comprehensive explanations than are currently available. There is also some interest in bringing in neuroscientific research into the court system. Currently, defendants can use neuroscience research in the court room to show evidence of brain damage and/or neurobiological abnormalities. A study by Farahany (2016) found that between the years of 2005 to 2012 the number of judicial opinions that used neuroscientific evidence in this way increased significantly, from 100 to 250-300 per year. There were two main reasons provided for including neuroscientific evidence: 1) *competency* – whether the defendant has the rational understanding to undergo a trial; and 2) *judging guilt* – including whether the conduct was involuntary, whether premeditation is possible for the defendant, attempts to mitigate sentencing, and claims about the developing brain for juveniles.

Aside from the above applications of neuroscience findings in adjudicating competency and guilt, there have also been recent attempts to integrate neuroscience into the court room even further. During a criminal trial, the mind of the defendant is under scrutiny to try and reveal whether they committed the crime voluntarily (*actus reus*) and with intention (*mens rea*) (Pardo and Patterson, 2013). Currently, the prosecution must rely on physical and circumstantial evidence, or eye-witness testimony. They may also use techniques such as cross-examining the defendant to extract the truth of the situation (Pardo and Patterson, 2013). However, it can sometimes be very difficult to prove guilt, especially if the evidence in a particular case is lacking. The allure of neuroscience is that it may be able to allow us insight into the mind of the offender that would otherwise be unavailable. For example, companies in the United States have been developing technology using fMRI to detect when a person is lying (Pardo & Patterson). Attempts to use these technologies in a court situation have failed due to issues with reliability, nevertheless there is a persistent effort to further develop this sort of technology so that it can one day be introduced. For example, the New Zealand Law Foundation has recently provided funding into a research project that is attempting to prove the efficacy of brain scanning technology for use in forensic settings (Barratt, 2016).

In my view it is encouraging that forensic psychology is increasingly turning to neuroscientific research to develop a deeper understanding of crime related phenomena and to guide legal and correctional practice. However, there are a number of major ethical issues that arise from importing neuropsychological evidence from the literature without consideration of its conceptual and theoretical validity. In forensic psychology, the evidence from the literature is often used to directly impact on peoples' lives. This is seen in therapeutic settings, where the RNR model is the most common framework used in correctional rehabilitation. The evidence concerning correlates of recidivism is also used for determining risk levels of offenders, and therefore any eligibility for parole. In a therapeutic setting, neuroscience evidence could be used to guide treatment targets and therapeutic interventions. Further, as previously discussed there is an impetus to bring more neuroscientific techniques into the courtroom setting – this means that this evidence could impact on legal judgement and sentencing. Therefore, it is essential that any research that is used in the forensic area is robust, reliable, and valid. This thesis will outline a number of important considerations that need to be acknowledged and addressed by researchers in the field to ensure that this is the case.

The second chapter will outline these considerations, including: 1) Issues of a metaphysical nature: What is the mind? How does it relate to the brain? 2) Issues of epistemology or explanation: What kinds of explanations are being posited? Are they reductionist, or multilevel? 3) Issues with the conceptualisation of cognitive processes: How valid are the concepts that we are using? Is there a consensus or a coherent taxonomy we can use? And 4) Technical issues associated with neuroscientific research: What are the constraints of neuroscientific technology? What kind of measurement issues are most prominent?

After detailing these considerations, the third and fourth chapters will examine these issues in current research associated with the forensic area, in order to explore how well they are being addressed, and whether or not they are even discussed. The third chapter will focus on the area of '*deception detection'*. As discussed, deception detection is a pressing topic considering the impetus to bring it into the court system. The fourth chapter will focus on the area of '*mind reading by brain reading'*. Similarly to deception detection, in recent years there has been some development of technologies that purport to 'read' the minds of defendants using neuroscientific techniques such as fMRI in order to ascertain knowledge

(Pardo & Patterson, 2013). These two chapters will evaluate the literature base in each area according to the considerations discussed in the first chapter. The fifth and final chapter will provide an overall summary of these considerations and their representation in the literature, and posit some potential solutions that researchers can utilise. This will include: making sure that research has a strong theoretical base, being clear about the type of explanatory claims the research is making, ensuring clarity and consistency about the definitions of the phenomena of interest and the methodology employed, and finally, acknowledging and adjusting for any conceptual and technical limitations.

CHAPTER 2: Considerations

The Metaphysical: the mind and the brain

When considering the conceptual issues in the area of neuroscience, it is first and foremost essential to look at the fundamental assumptions about the mind and brain. These assumptions guide both the initial direction of research and the interpretation of results. It is also important to consider what it is we are trying to explain, and how we go about explaining it.

Reductionism

In neuroscience, it has become increasingly common to hold a 'neuro-reductionist' position, wherein the mind is reduced to the brain: the mind is the brain, and all mental activities and attributes can be explained somehow by utilising information about brain structures and processes (Pardo and Patterson, 2013). Reductionism is considered necessary, because the use of this philosophical position allows one to ignore other levels of explanation (such as the psychological) and explain behaviour solely in terms of brain activity, allowing for simpler explanations. However, the reduction of the mind to the brain is conceptually problematic for many reasons.

Pardo and Patterson (2013) outline a number of these issues with the neuro-reductionist approach. First is the *mereological fallacy*. This is the idea that we cannot attribute to a part what is only attributable to the whole (Pardo & Patterson, 2013). Reductionism relies on the belief that the mind and brain are one, but this ignores other levels of explanation, and cannot account for aspects of psychological functioning such as social interaction, knowledge, intentionality, and decision making. For example, a concept such as 'love' cannot be reduced to a response to single a simple stimuli, as it is not a single enduring state. On the contrary, it is ever-changing, based on a person's interactions with others, and can include many different emotions – happiness, jealousy, anxiety, excitement, guilt**.**

Despite this, several studies have attempted to find the neural correlates of love using fMRI techniques (Cheng et al., 2010; Fisher, Aron and Brown, 2005). Ultimately, the issue here is whether the mind is viewed as a thing that can be located and quantified, or whether instead we are looking at how a *person* acts and interacts with others and the outside world. While neural activity is an important part of how we understand the person, arguably it cannot be considered a full explanation in and of itself.

One reason this is the case is the issue of *criterial* versus *inductive* evidence. Criterial evidence is that which is constitutive of a phenomenon, and inductive evidence is that which is correlated with a phenomenon. For a reductionist account to be conceptually valid, the physical state of the brain would have to be criterial evidence of psychological phenomena. However, this is not the case: brain states are not constitutive of psychological phenomena; they are a measurement rather than a direct measure (Pardo and Patterson, 2013). The brain state itself is not enough to explain a phenomenon: when a person perceives their environment, it is explained in terms of what they see, think, and feel – not what their brain is doing. Further, inducing brain states would not necessarily recreate the phenomenon – while it is possible to induce sensory experiences there is no evidence that it is possible to recreate more complicated phenomena, such as beliefs or intentions. Therefore, the removal of other levels of explanation is conceptually problematic.

Another issue is that of *knowledge*. In legal settings, ethical and legal judgements depend on finding out what a person did or did not know, or were capable of knowing at the time of the crime. Neuroscientific explanations purport to ascribe a location to knowledge – that it is stored in a certain area of the brain, and that this can location can be discovered through the use of neuroscientific techniques such as fMRI or PET scans. However, as Pardo and Patterson (2013) explain, knowledge should not be thought of as simply something one has; rather it is something that is manifested through behaviour. Having the neural correlates of a psychological phenomenon is not enough to show knowledge:

> "Knowing is an ability – the ability, for example, to answer a question correctly. The measure of the truth of your answer is not found in a neural state of your brain. Whether you know the answer to the question is shown, among other ways, by what you sincerely say in response to the question." (Pardo & Patterson, 2013, p. 21)

Another reason why it is important to have clarity concerning the concept of the mind assumed by neuroscience researchers concerns agency, or one's ability to engage in intentional action. As Runyan (2013) points out, there are two ways to explain agency: *reductive*, where our experiences can comprehensively be explained by subpersonal (neural) causes; and *emergentist*, where our experiences can be explained in terms of what people do and cause themselves. The reductive position aligns itself with the neuro-reductionist model, and therefore is supported by those who believe that the mind is completely constituted by the brain. The most well-known experiments in this area are the Libet-style experiments, where the conscious awareness of an intention to act was compared with concomitant neural signals, finding that there was a 'Readiness Potential' that occurred prior to the participant's own awareness of their will to move (Libet, 1983). This provided convincing evidence to many that 'will' is illusory and superfluous to human action. There are serious moral and ethical implications to taking this point of view, particularly in the forensic area. If there is no such thing as will, and all actions are predetermined, then this would call into question predominant ideas about criminality. For instance, whether the person is responsible for the crime, and subsequently the extent to which they should be punished for actions outside of their control.

Embodied Cognition

One of the more prevalent theoretical orientations in contemporary psychological science is the computation theory of mind (CTM). This approach describes the brain as similar to a computer, where it receives inputs, compares these inputs to representations stored in the brain, and then provides outputs based on what these representations convey. Here, as in the reductionist approach, the mind is seen as indistinguishable from the brain, as all phenomenological experiences are caused by underlying brain structures. In this way of thinking, all behaviour should ultimately be predictable, as long as the tools used to analyse the brain are accurate and reliable. Traditionally, cognitive science explains mental states as representing features of the world, and mental processes transform and manipulate these inner representations. In this way, physical brain states are like ‘hardware’ and mental processes are like ‘software’ (Drayson, 2009).

Not all researchers subscribe to this theoretical orientation. An emerging view of psychological functioning that has become more popular recently is that of *embodied cognition*. This espouses the idea that cognition extends outside of the brain, into the body, and into the direct environment. As Drayson (2009) explains, embodied cognition embeds the biological brain within the body, its natural setting, and emphasises the brain and body’s interaction with the environment. Without this conception, the study of cognition is “radically disembodied” (Drayson, 2009, p. 5). The idea of *enactivism* goes a step further, not just placing the brain within the context of the body, but linking perception and action as connected, intertwined concepts. Here, cognition is viewed as a direct result of the linkages between the sensory and motor systems. This removes the (arguably arbitrary) boundary between the internal and the external, and extends cognition beyond the brain and into the world. The argument here is that our brains have evolved to control and respond to action, using complex causal networks involving the brain, the body, and the environment. While brain-based inner representation is downplayed in an embodied explanation, Drayson argues that it could still exist in cases of offline cognition, such as mental addition. This type of approach could help to explain emergent properties at higher levels of analysis, or why neuro-reductionist approaches can only offer a partial explanation.

Neural Network Approach

The Neural Network Approach is a comprehensive interactionist model developed by Anderson (2014). It was derived from embodied cognition rather than CTM, as Anderson argues that the evidence points more strongly to an interactionist model rather than a modular one. Underlying this approach is the idea of *neural reuse*, wherein neural structures can be used and reused for a wide range of purposes. While this is classically seen in examples of neural plasticity and neural reorganisation after brain injury, Anderson contends that the most compelling evidence for neural reuse is the diverse and continually evolving tasks (due to, for example, technological advances) that we learn to complete. Rather than developing new and specialised neural structures for new tasks the brain is adept at reusing existing structures. For example, the invention of the automobile did not result in new brain structures being formed in order to learn how to drive. Rather, our brain reused the neurons that were already evolved for tasks such as navigation and motor control. More recently, there has been an explosion in the number and types of technology that we are exposed to, including computers, cell phones and smart phones, and even virtual reality. Again, there are no new developments in brain structure, but reuse of the structures already in place. Anderson also argues that this is why we often see widespread activation across the brain: any new developments in cognition would ideally use the most useful existing elements, regardless of relative locations in the brain. On the other hand, more archaic features, such as vision and motor cognition, would be more specialised (as is evident in the occipital and motor cortices). The empirical evidence appears to show that neural structures are used for multiple purposes rather than a single function. This is referred to as *functional differentiation*. Furthermore, Anderson proposes that it is essential to place this viewpoint within the embodied cognition framework. The reason behind this, he argues, is that the brain evolved to control action, and that the coupling between perception and action is critically important for understanding how the brain evolved. Neural reuse as an explanation of the architecture of the brain can account for how these perception-action feedback loops can result in functional differentiation. Therefore, CTM represents a fundamental misunderstanding of the nature of the brain.

Uttal (2011) provides a detailed look at the empirical evidence in cognitive neuroscience through a number of different phenomena including sensation, perception, emotion and affect, learning and memory, attention, and consciousness. Throughout these fields it becomes clear that there is no real consensus about the roles of specific brain regions, with high amounts of variability within subjects and between studies. For example, he comes to the conclusion that:

> "Traditional ideas of special roles for restricted areas of the brain must be re-evaluated. There is no evidence of any unique role for any brain region in perceptual encoding or in any other domain of cognitive neuroscience." (Uttal, 2011, p. 138)

Even processes such as face recognition, which seem to have a large amount of evidence pointing to the role of the fusiform gyrus, continue to generate controversy and inconsistency among researchers about the domain-specificity of the fusiform face area (see Kanwisher & Yovel, 2006; McGugin, Newton, Gore & Gauthier, 2014).

Considering that CTM has guided the majority of cognitive neuroscientific research and continues to be the predominant theoretical model that is employed in research, it is important to consider whether it is conceptually and theoretically valid.

Explanation

Aside from determining the theoretical orientation underpinning how we conceptualise the mind, it is also important to examine closely how particular concepts and processes are explained in psychology. Ward and Wilshire (2018) describe the most important steps for developing explanation in the area of forensic psychology, including: determining the focus of inquiry; distinguishing between different forms of explanation; and differentiating between multilevel and reductionist explanation.

Focus of Inquiry

When considering research in any area it is important to narrow down the focus of inquiry, so that a clear description of the phenomena of interest is evident. As Ward et al. (2018) explain, in forensic psychology inquiry is often based on normative categories of offending behaviour, rather than scientifically valid and theoretically directed explanatory targets. On the other hand, difficulties may arise when attempting to apply research from neuroscience to the forensic area because of the issues in how explanatory targets are defined differently. This is explored below when discussing problems with concept validity in psychology.

Etiology vs. Composition

One important distinction in explanation is whether the type of explanation is etiological or compositional. An etiological explanation is one where a set of *causal* processes (or mechanisms) result in a subsequent downstream effect, while a compositional explanation is one where the constituent components of a single mechanism are investigated (Ward & Wilshire, 2018). In order to provide an etiological explanation, causal mechanisms need to be identified. Mechanisms produce, underlie or maintain phenomena, reliably and predictably generating changes (Ward and Wilshire, 2016). On the other hand, a compositional explanation does not require the identification of *causal* mechanisms. Rather, it is interested in the mechanisms that occur as *part of* the phenomenon. For example, a compositional explanation would look at the neural processes that occur during a particular behaviour, rather than looking at what causes that behaviour.

In forensic psychology we are particularly interested in looking at the causes of offending, whereas empirical studies often provide or inform compositional explanations (Ward et al., 2018). Further, research in neuroscience also often takes a compositional approach due to the issues in establishing causality. While compositional research is valuable and an essential addition to our understanding of phenomena, it is nevertheless important to be aware of this distinction when conducting research, and when transposing research from the neuroscience area to forensic psychology. As Ward et al. (2018, p. 9) argue:

> The challenges associated with drawing inferences from neuroscience findings can be understood in terms of the distinction between compositional and etiological explanation… It is relatively easy to make the assumption that a lower-level description of a phenomenon somehow speaks to its cause. However, this is usually not the case; neuroscience findings can certainly help us to develop a compositional explanation of a phenomenon, but they rarely provide an etiological one.

Multi-level explanations

Furthermore, another important distinction to make is whether the explanation is reductionist or multilevel. As mentioned previously, in neuroscience the reductionist position is common as this allows the researcher to use the implicated neural structures or functions as the primary level of explanation for the behaviour in question. However, this ignores the other levels of analysis that are arguably just as important as the molecular/brain level in providing a comprehensive explanation. Craver (2007) refers to this as an explanatory mosaic, a theory that takes the necessary evidence from different levels of human functioning to create an explanation that more fully encompasses the phenomenon.

Thagard (2016) proposes that the best view is one that uses *multilevel emergence*, showing how different levels of explanation are connected. An emergent property is one that is possessed by the whole, but not the parts. The phenomenon in question cannot be seen at more reductive levels, but is produced by the interaction of these levels. For example, while neural firing and biochemical reactions in the brain are necessary for cognition, they cannot fully explain why we behave in a certain way, for instance, committing theft or assault. It is necessary in these cases to look at the phenomenological and social levels in order to explain, firstly what the behaviour is, and secondly the reasons why the action occurred. Patterns of brain activation or structural differences may provide some insight into causality, but they do not by themselves provide a sufficient explanation. For example, while a table may be made up of molecules, to explain the concept of the table we need more information (i.e., its function, parts, materials, etc.); explanation is not meaningful solely using the molecular level. Rather, all levels of explanation add to our understanding of what a table is (Thagard, 2017).

Additionally, it is not simply that the lower levels impact on the higher levels, as all levels can variously impact on others. For example, taking the case of theft, it could have a number of causes: 1) molecular – methamphetamine use affects neurotransmitter release, which causes impulsivity and reckless disregard for others; 2) neural – abnormal patterns of activation lower inhibition of behaviour and executive decision making, and increase impulsivity; 3) psychological – personality or other psychological predispositions towards offending behaviour; or 4) social – antisocial peers, or difficult life circumstances (i.e., financial debts). While the higher levels (e.g., the social) rely on the lower levels to occur (as in, there could be no social interaction without the accompanying neural response), the higher levels can also affect the lower levels (for example, receiving an insult would create a corresponding neural pattern and biochemical response).

Explanatory pluralism

Additionally, as Thagard (2017) argues, different phenomena may be best explained using either a certain type of explanation, or a combination of explanatory methods in order to encapsulate the phenomena. He states that there are generally four types of explanation: Narrative, Deductive, Mechanistic, and Eliminative. A *narrative* explanation is the traditional type of explanation: a story that answers the question of why something happened. Though it is more commonly used in storytelling, it can also be used in science to explain certain processes. For example, processes such as the big bang and evolution are often described with a narrative structure. In social science, such as economics or history, a narrative structure is also often employed (Thagard, 2017). Narrative explanations are satisfying, as they explain a process in a way that appeals to people. However, causality is not emphasised here. Narrative explanations are useful where causality is not clear; where the process is understood to occur but we do not know why. However, if possible it is best to improve the narrative explanation by identifying the mechanisms that underlie the process.

The second type of explanation is *deductive*. Here, phenomena are logically deduced from general laws and a set of initial conditions; explanation is essentially a deductive argument. A conclusion is deduced from the relevant laws, with no consideration of the underlying mechanisms. There are a number of advantages of using a deductive explanation: they are logically rigorous; there is a clear connection between the phenomenon and explanation; often the best explanation is inferred; and these models are often useful for prediction (Thagard, 2017). Because of this, deductive explanation can be a good way to identify potential causal factors in behaviour. However, these explanations are not enough to establish causality as they are more focused on correlations rather than causal mechanisms, essentially describing patterns of phenomena. An example in the forensic area would be risk factors. We have enough deductive evidence to establish some relationship between a risk factor and offending behaviour, but this does not provide a complete explanation. For example, we know that there is some relationship between substance use and offending, and that it predicts future offending, but this does not provide a complete explanation of the underlying mechanisms and explanation of the cause of the offending.

The third type of explanation is *mechanistic*. As previously discussed, a mechanistic explanation is one that focuses on the causal mechanisms that generate or underpin a phenomenon. While this is the most common type of explanation in psychology, and one that is considered superior, Thagard warns against generalising and prioritising a mechanistic explanation to a general theory of explanation because other types of explanation, particularly deductive, can provide valuable explanation when mechanisms are not known.

Lastly, Thagard (2017) describes *eliminative* explanation. This is where the purpose of an explanation is to eliminate a previous explanation for a phenomenon. Thagard gives the example of mental illness replacing demonic possession as an explanation for particular abnormal behaviours. This type of explanation is used by eliminative reductionists, who seek to eliminate the mind in favour of describing behaviour as purely biological. However, Thagard (2017) argues that eliminative reductionism is essentially a failed enterprise due to the fact that the mind as we experience it is a real phenomenon (i.e., the phenomenological or experiential level of analysis), and that neural explanations support and substantiate this phenomenon rather than replace it. He uses the example of a table: just because we know that a table is made up of molecules it does not eliminate the table itself, rather it provides more evidence for its existence.

Therefore, it is important to consider a number of different types of explanation in theory. This is described as *explanatory pluralism* by Mantzavinos (2016). As he explains, the idea that one type of explanatory model can account for all explanatory activity is problematic. For example, the causal mechanistic model is often sought after as the key to understanding human behaviour. However, while this model may be able to encompass certain aspects of behaviour, it cannot encompass it fully. Rather than trying to explain phenomena with the ‘best’ type of explanation, the goal should be to capture and integrate a multitude of explanatory models.

Additionally, Mantzavinos (2016) points out that explanation is socially entrenched, and that knowledge is produced within certain cultural and institutional constraints and according to normative standards. For example, what Mantazavinos refers to as constitutive rules determine 1) what counts as an explanandum, in that the phenomena or behaviour we are studying will be bound by our current knowledge and ability; 2) what must be taken as given, or what research has already occurred which informs our explanation; and 3) the metaphysical presuppositions, in this case our presuppositions about the mind and the brain, will orientate our explanation in certain ways. Therefore, explanation will naturally evolve and expand over time. By limiting ourselves to certain explanatory models we limit our ability to gain further knowledge. Explanatory pluralism is a way to expand the way in which we perform research.

In summary, it is important to 1) be clear about whether you are using an etiological or compositional explanation; 2) ensure that explanation is multi-level rather than reductionist; and 3) consider the contribution of different types of explanation in developing a theory.

Methodology

Finally, it is also important to examine the practical issues inherent in neuroscience research.

Validity and Reliability

Sullivan (2015) describes reliability and validity as normative constraints on the experimental process. They require researchers to achieve a delicate balance in order to create research that tells us something meaningful about the phenomena in question. Research that is high in reliability is replicable, and it is predictable. However, this often requires a simplification and idealisation of the behaviour as it would occur in the real world, into an experimental design that can be easily systemised. On the other hand, experiments higher in validity will have a design that is complex and nuanced enough that the results can be reasonably extrapolated to the real world.

Ecological validity refers to the similarity between the experiment and the real world, in terms of the context, stimuli, and responses (Sullivan, 2015). Although it is generally impossible to have an experiment that is identical to the real world (due to the importance of informed consent and ethics in psychology), it is generally believed that the closer the experiment is to the real-life situation that would ordinarily elicit the desired response, the more accurate and theoretically valid the experiment is. Neuroscience research has a particular problem with ecological validity. While experiments in other fields, such as social psychology, can more closely approximate a given social situation, research in neuroscience relies on data that needs to be collected under controlled circumstances. In fMRI research, this requires the participant to be lying completely still in a scanner, often responding using a button press. This allows rich data to be collected about the processes occurring in the brain, but calls into question whether this can adequately generalise to real-life. Additionally, because of these restrictions, the tasks themselves have to be simplified, and the participant is very aware that they are in fact completing a task rather than providing a natural response to a situation. As Uttal (2011) explains, often neuroscience/cognitive research will focus on one particular cognitive theme (such as visual attention) in isolation. This ignores other intrinsic features of cognition and perception. Further, there is a lack of control over the participants' attention and motivation in the task (Uttal, 2011). It is possible that there could be differences in brain activation when compared to real-life situations, particularly in the frontal lobe (e.g., attentional processes) and limbic system (e.g., due to emotional reactions).

Construct Validity

Another important challenge in this area revolves around *construct validity*. This is not just an issue in neuroscience, but throughout psychology as a discipline. As Uttal (2011) explains, defining cognitive processes imprecisely leads to a large variability in evidence about the process during different experiments. This problem is more severe the more abstract a phenomenon is. For example, a phenomenon such as visual attention is clearly defined; it is easily explored by following eye movements and reaction times. On the other hand, a phenomenon such as creativity is much more nebulous and difficult to define. The problem of definitional vagueness leads to different laboratories designing different experimental protocols to ostensibly explore the same phenomenon, resulting in a number of different answers about the factors that contribute to this phenomenon. This then leads to comparisons between studies that can be either misleading or confusing.

Sullivan (2008) expands on this problem, arguing that there has been a lack of attention paid to the differences in the *experimental protocols* evident in different laboratories. Subtle (or large) differences in the tasks used by these laboratories to investigate the same phenomenon may lead to differences in the effects seen, and in neuropsychology, the areas of the brain that are implicated. Sullivan argues that you therefore cannot directly compare the data obtained from another laboratory unless the tasks used are identical. Additionally, it is problematic to investigate the broad concept using these different experimental protocols.

Sullivan (2015) also makes the point that when conducting research, the effect that is being studied should be an "actual instance" of the phenomenon. For example, if you design an experiment to look at the phenomenon of visual attention, the experiment should actually be measuring visual attention. This seems to be obvious on the surface, but upon closer inspection it is not always clear that the effect lines up with the phenomenon. Sullivan (2010) gives the example of the Morris water maze: while it is commonly known as a test of spatial memory, it has also been interpreted as spatial learning, spatial memory, spatial navigation, and spatial cognition. While these terms are related, they do capture separate processes, and therefore it is important to be accurate about what exactly is being investigated. Similarly, often cognitive processes are being explained at a 'coarse' rather than a 'fine' grain (Sullivan, 2010). For example, an experimental protocol such as the aforementioned water maze may attempt to explain 'spatial memory', which is a somewhat vague construct that contains a number of different processes and neural systems. The explanatory target 'spatial memory' is not specific enough to provide a coherent explanation.

When we conduct experiments based on behaviour, we infer cognitive constructs from neural observations. This is where the problems occur. The addition of indirect measurement (e.g., fMRI) adds credibility and affirmation to these constructs, as if they can be directly measured in a physical sense.

Cognitive Ontology

These issues with construct validity emphasise the need for an *ontology of cognitive science*, as explicated by Poldrack (2010), one that classifies mental processes in a way that correctly identifies "natural kinds", and as such can uncover the causal processes occurring in cognition.

As Sullivan (2016) explains, it is not currently possible to integrate psychological and neuroscientific explanations to create a unified science of cognition. This is because there are few stable constructs in the neuroscience of cognition. While neuroscience focuses on providing mechanistic explanations, cognitive science focuses on deductive explanation provided by functional analysis. In order to create a unified science, what Sullivan calls a ‘bridge law’ must be established. This means that the terms (referents) in one theory, the theory that is being reduced, must be equivalent to the terms in the reducing theory. The example provided here is one in physics: the temperature of gas (theory being reduced) is equivalent to the mean kinetic energy of the molecules (reducing theory). Therefore, in order to unify two theories, you must be explaining the same thing. However, in cognitive psychology and neuroscience, there are *no stable referents* – terms can be defined very differently by different research labs. Stabilisation requires coordination across the field in the following areas: 1) in how to generally define the terms; 2) in determining the best experimental paradigms for studying this concept; and 3) the conditions under which two experimental paradigms can be said to measure the same concept (Sullivan, 2016). While some tasks can be thought of as a gold standard within a particular area of study, in general there is no agreement throughout the discipline as to which experimental paradigms should be used to investigate a particular construct, and there is freedom to design tasks that each particular researcher or group decides will be most appropriate for their investigation. While this may be useful for initial experimental exploration into a given construct, it poses a problem for this idea of construct stability and thus a unified science of cognition. If the tasks, while seemingly searching to explain the same construct, are actually targeting different aspects of cognition, then the construct itself becomes unclear and unstable.

However, Sullivan emphasises the importance of different methods and perspectives, and proposes that we should keep intact the perspectival pluralism in cognitive neuroscience, as a path towards further explanatory integration. An example currently being developed in the field of psychopathology is the Research Domain Criteria (RDoC). The aim here is to improve the current classification of mental disorders from one that is based on symptom clusters (as in the DSM-5) to one that is biologically based and empirically validated (Insel et al., 2010). The goal is to develop a research framework for scientists in the field that will guide experimentation in a way that can identify the specific behavioural and cognitive disruptions in those with mental disorders (Sullivan, 2016). While the RDoC focuses on mental disorders, this kind of framework would be valuable within the wider discipline of psychology.

Poldrack et al. (2011) argue for the need for a ‘cognitive atlas’; a formal knowledge base which would accurately capture a wide range of cognitive processes. One reason for this is the ambiguous terminology that is used in psychology. For example, the concept of working memory has a number of distinct definitions in use throughout the field, including holding information in memory, manipulating information in memory, and memory for varying aspects of a task (Poldrack et al., 2011). Additionally, they argue that there has been a conflation of tasks and constructs, wherein the task is assumed to be a full representation of the construct. However, it is possible that a task may only tap in to certain aspects of a construct, or it may tap into several constructs, depending on the structure of the task. Two common examples are that the Stroop task is also often referred to in the literature as the ‘inhibition task’ and the N-back task is referred to as the ‘N-back working memory task’. While this conflation may be in part due to conciseness on the part of the author, it also represents a worrying trend where the construct itself begins to become defined by the tasks that are most often used to measure it.

Technical Issues

There are also a number of technical considerations that need to be addressed in neuroscience. The majority of current research utilises functional magnetic resonance imaging (fMRI), with a significant proportion of the remaining studies using electroencephalography (EEG) and a small minority using techniques such as transcranial magnetic stimulation (TMS) or Positron Emission Tomography (PET). Thus, it is prudent to focus on fMRI as the most common method, and one that is often considered the gold standard for neuroimaging. As a preface to discussing these considerations, a brief overview of the method is as follows: fMRI uses blood-oxygen levels in the brain (referred to as the blood oxygen level dependent, or BOLD, signal) to indirectly measure neural activity. It is assumed that neural activity is correlated with the BOLD signal, with the logic that increased activity increases energy requirements, provided by oxygenated blood. Once the magnetic field detects the polarity of ions in the bloodstream, statistical testing (often using a subtraction method, subtracting the 'resting' state of the brain to the 'active' state) reveals areas that are differentially activated in the applicable experimental paradigm. While this is an incredibly useful imaging technique, and one that has revolutionised and invigorated the neuroscientific community, there are a number of concerns which need to be considered when interpreting results. These include 1) problems with reverse inference; 2) specificity of neural activation; 3) indirectness of measurement; 4) noise; 5) subtraction issues; and 6) the seductiveness of brain images.

As Poldrack (2006) outlines, there is an overabundance of reverse inference in fMRI studies. Reverse inference is where the activation of a particular region of the brain is detected, and the relationship to a cognitive function is inferred backwards from this. Often this occurs when an fMRI experiment shows activity in unpredicted or unexpected regions; it is assumed that this activation is related to the cognitive process that the experiment was designed to elicit. Poldrack notes that this is the logical fallacy of affirming the consequent: because the region was activated during a particular task, it must therefore mean that it is necessary for the task. However, because brain regions have low selectivity for activation – there is a lot of crossover in activation between regions – this is not by itself good evidence to say that this particular region is involved in this cognitive process. While this type of experiment is useful for exploratory research, it should be best used to suggest novel hypotheses that can then be tested in subsequent experiments. This would also preferably be used in combination with data from other levels of explanation, such as the behavioural level (Poldrack, 2006).

Uttal (2011) describes this issue with low selectivity as the 'one-to-many' and 'many-to-one' problems. The one-to-many problem is the idea that for each area of the brain there are a number of different cognitive processes that activate it. It is problematic to assign a specific meaning to a certain area because it showed activation on a brain scan without eliminating other mental processes that may be going on. The many-to-one problem is the idea that cognitive processes are often represented by a number of brain areas. This may also vary depending on individual brain differences. This can also be explained as the difference between a 'sign' and a 'code' (Utall, 2011). A sign is simply a correlate of brain activity; it tells you where the brain is active during a certain task. In contrast, a code is the 'psychoneural equivalent' of the cognitive process; it is both necessary and sufficient for the process to occur. By using reverse inference we can discover many signs of cognition, but to discover a code requires much more rigorous, theory-directed experimenting.

Further technical issues arise from the method of experimentation itself. While fMRI has provided a number of valuable insights into the workings of the brain, there are a few technical challenges that impact on its usefulness. The first is the indirectness of the measurement. As explained earlier, the data that is collected relates to the BOLD signal, rather than neural signals themselves. The BOLD signal is assumed to be correlated with neural signals, as an increase in activity would naturally require an increase in energy, which is obtained from the bloodstream. However, this is still an indirect measurement, and is quite a 'large grain' view of brain activity, rather than the 'small grain' of neural firing. It is important to keep in mind that while it appears that fMRI is taking a direct measurement, this is not entirely correct. A further issue that arises here is that metabolic processes take much longer than neural signals to occur (Uttal, 2011). The solution here is again to make sure that your experiment is conceptually clear, and theory-driven.

There is also an issue with the so-called ‘seductiveness’ of the brain images (Uttal, 2011). This particularly applies in a court situation, where the jury may be asked to view images of brain scans. However, it also applies to the general public’s view of neuropsychological imaging. It appears that laypersons attribute more authority to brain images than they are worth. This occurs because they do not have the knowledge of the fMRI process nor the limitations that is needed in order to successfully evaluate neuropsychological results. As Sullivan (2015) explains, there is no appreciation that the images are outcomes of statistical interpretation and manipulation of data, instead of being direct outcomes of the fMRI process itself. Fortunately, research has shown that those with this knowledge do not over-value the brain images (Weisburg et al., 2008). Ultimately it is important to keep this issue in mind when considering using neuropsychological research in the forensic area, especially as it is communicated to novices.

Further, another consideration when using fMRI results is the subtraction issue (Uttal, 2011). To view activity in a specific task, the control condition (activity at ‘rest’) is subtracted from the experimental condition. This has the effect of ‘zero-ing’ out the activity. However, there are a number of problems with this. First, just because the activity was present in both conditions, does not mean that it is not necessary for both. Second, because fMRI takes such a course grain look at brain regions, due to it being a secondary measure of blood flow rather than directly measuring neural firing, it is not clear whether the activity in the same region is actually the same – the underlying neural activity may change from control to experimental condition, but it is not possible to see these small differences. Third, this assumes that the brain is not active during the control condition.

While fMRI is currently the most popular method of experimentation in neuropsychology, it is possible that new methods, or using a combination of methods may help to minimise the preceding technical problems. One relatively new technique that shows some promise is transcranial magnetic stimulation (TMS). TMS works by disrupting the neuronal firing in a certain area of the brain using a magnetic pulse. This allows researchers to observe the necessity of particular areas of the brain in certain tasks, which allows for stronger claims about causality.

Summary

Overall, the considerations when bringing in neuropsychological evidence into the forensic psychological area are as follows: 1) the attention paid to the theoretical underpinnings of research, including how the mind is conceptualised; 2) the kinds of explanations that are being used, and whether there is one kind of explanation or explanatory pluralism; 3) whether the research focuses on one level of explanation, such as the neurobiological level, or multiple levels; 4) whether the concepts being used are valid and defined consistently throughout the field; 5) the methodology employed, and how consistent this is between laboratories; and 6) whether the technical limitations of the methodology are adjusted for or acknowledged. The next chapter will examine the literature in two areas that are relevant to forensic psychology: deception and mind-reading by brain-reading. I will assess how well the literature in these two areas has taken these considerations into account.

CHAPTER 3: Deception Case Studies

Deception is a complex concept. It can be defined as "the act of causing someone to accept as true or valid what is false or invalid" (Merriam-Webster, 2018). There are two aspects that need to be considered here: the *intent* of the deceiver and their *actions*. For intent, the keyword here is *causing*: it must be a deliberate act. Lies based on misinformation or that occur by accident cannot be considered deception. On the other hand, deception does *not need to include lies*: a person can tell the truth in a way that is intended to deceive another. In terms of the actions involved, Lippard (1988) identifies a number of behaviours, including lies, half-truths/distortions, exaggerations, withholding information, cheating, stealing, and hiding behaviour. It is important to have a clear definition of the concept being studied, as vagueness and imprecision can undermine the conceptual clarity and coherence of the results.

In this chapter I will analyse the literature based on the following questions: 1) how the mind is conceptualised; 2) what type of explanation is being used; 3) how the concept of deception is defined and operationalised; and 4) the methodological and conceptual issues evident and how they are accounted for.

Literature Analysis

In a study by Davatzikos et al. (2005), participants were given two playing cards, and were instructed to deny possessing one card and acknowledge possessing the other. The experiment contained five stimulus classes: lie (one card possessed by the participant), truth (the other card possessed by the participant), a recurrent distracter, varied distracters (the remainder of the playing cards), and a null stimulus (the back of a card). Response was obtained through pressing the left button (yes) or the right button (no) to confirm whether the card was in their possession. The results showed that there were two main regions of activity:

the right prefrontal cortex, including the inferior and superior frontal gyri; and the bilateral posterior cortex, including the superior temporal and inferior parietal gyri. Activity was determined using a pattern classification method. In this study, the conceptualisation of the mind is not addressed, but the methodology suggests a CTM orientation given the search for significant regions of activity. The type of explanation here is purely compositional as there are no causal claims being made. Additionally, it is mechanistic in that the research is designed to discover the neural regions that underlie deception. The concept of deception is not explicitly defined here. However, it is operationalised as lying about possession of an object: the intent is to cause the experimenter to believe a falsehood, and the behaviour could be defined as either general lying behaviour or withholding information. There is some discussion of the practical limitations of this method, including the fact that it would be time consuming to collect baseline data for each individual, and that their pattern classification method does not identify important brain regions in advance and therefore may include brain areas that are just noise (irrelevant).

A second study examining the concept of deception was carried out by Volz et al. (2015). Here, the method was a sender-receiver game, which included two players – one being the sender and the other the receiver. The sender was the one being scanned for neural activity. The idea of the game is that the receiver has two options to pick from, red and blue, and the sender is to send the receiver a message indicating which colour would be more profitable for the receiver (as the game had perceived monetary consequences). The sender then indicates which colour they expect the receiver to pick. There are three outcomes here: "Sophisticated deception" in which the sender has sent a true message but expects the receiver not to follow it (as they would believe that the receiver is suspicious of the sender's intentions); "True message", where they send the true message and expect the receiver to follow it; and "Simple deception" where the sender sent the false message and expected the receiver to follow it. The results showed that a number of brain areas were implicated in this deception task. For both the simple and sophisticated deception, the areas included the right temperoparietal junction, the superior temporal gyrus, the precuneus extending into the retrosplenial cortex, the bilateral cuneus, and the right superior frontal gyrus. For the simple deception, areas also included the dorsal anterior cingulate cortex and the right anterior frontal gyrus. For sophisticated deception, areas also included the left cuneus and the right anterior frontal gyrus. Again, the theoretical orientation was not explicitly discussed; however, while the data analysis was based upon ROIs, the authors acknowledged the importance of networks of neural activity, rather than simply these regions of interest (ROIs). Despite this, the orientation is still modular and aligns with CTM. This research was also compositional, and mechanistic. Deception is defined as “a deliberate act that is intended to foster in another person a belief or understanding which the deceiver considers false” (Volz et al., 2015, p. 1). The researchers then broke this down into what they term ‘simple’ and ‘sophisticated’ deception. For both types of deception, the intent is to cause the receiver to choose the least profitable option. For simple deception the behaviour includes lying, but for sophisticated deception there is no deceptive behaviour as the action is carried out by simply telling the truth. This study did not discuss any conceptual or technical limitations.

In a study by Ganis et al. (2003), participants were asked to write down the details about their most memorable work experience and most memorable vacation experience, and were also asked to generate, rehearse, and memorise a false scenario (related to one of the true scenarios). There were two lie conditions: the Memorised-Scenario (MS) lie condition, where the participants were asked to generate a lie based on the false scenario; and the Spontaneous-Isolated (SI) lie condition, where participants were asked to lie about the real scenario. There was also a truth condition where participants were asked to answer truthfully based on the information that they gave originally for both the MS and SI conditions.

Answers to questions were given with a binary yes/no button push or verbal single word answers. Results for both conditions indicated activity in the bilateral middle frontal gyrus, bilateral fusiform/parahippocampal gyrus, the right precuneus, and the left cerebellum. Results for the MS condition separately included the right anterior middle frontal gyrus. For the SI condition separately, areas included the anterior cingulate, the left precentral gyrus, and the right precentral/postcentral gyrus. Similar to the previous research, no conceptual orientation is discussed. Additionally, as this experiment is looking for neural correlates it can be classified as compositional and mechanistic. The definition used in this research is "deception occurs when one person attempts to convince another to accept as correct what the prevaricator believes is incorrect [typically in order to gain a benefit or avoid punishment]" (Ganis, 2003, p. 830). Interestingly, this definition includes potential motivations behind the deception. The authors also separate deception into two subcategories: spontaneous and memorised lies. The way in which deception is operationalised is based on episodic memory, rather than semantic. The intent here is to convince the experimenter of a falsehood, either based on a fictional scenario (MS) or a real one (SI). The actions for the MS lie condition are therefore both generating a lie, and then lying about this lie. The actions for the SI condition are simple lies. Interestingly, the researchers considered the MS truth condition to be equivalent to the SI truth condition, even though the scenario was fabricated. It is therefore possible that the comparisons are not entirely valid. This study includes substantial discussion on the limitations of this method of deception detection. Firstly, they highlight that the use of group analyses is not necessarily transferrable to a single participant situation. Second, they discuss the problem that their participants would not necessarily experience any guilt or emotional investment in the deceptive act, as they might in a real-life situation.

Another study by Ganis et al. (2011) used a different method to look at deception. The paradigm here was a concealed information test. Participants were required to lie about knowing their birth date. There were three conditions, the first being the no knowledge (control) condition, where the probe was an irrelevant date. Secondly, they had the concealed knowledge condition, where the probe was the participant's birth date, and the participant was instructed to lie about whether they knew this date. Lastly was the countermeasure condition, which was similar to the concealed knowledge condition, but participants were also instructed to perform a countermeasure (moving the left index finger, the middle left finger, or the left toe). The countermeasure condition was included to discover whether moving slightly could obscure the results (as it would cause additional brain activation). Results for the concealed knowledge condition, in comparison to the no knowledge condition showed activation in the cingulate gyrus, superior frontal gyrus, medial frontal gyrus, inferior frontal gyrus, insula, inferior parietal lobule, supramarginal gyrus, thalamus, caudate nucleus and lenticular nucleus. Results for the countermeasure condition found that there was increased activation in the primary motor cortex (as expected), but also found that the differential activation between the concealed knowledge and no knowledge conditions became much smaller. There was also a large increase in the number of false negatives – based on activity patterns, most of the deceptive trials would be classified as honest. In step with the previous research discussed, conceptualisation of the mind is not discussed, and this study follows the typical pattern of compositional, mechanistic, ROI-based investigation. No specific definition of deception was provided for this study. The experimental protocol was changed considerably from Ganis et al. (2003), as it used semantic rather than episodic knowledge. Therefore, the intent here is to withhold information from the experimenter, and the behaviour is a simple lie. Interestingly, this study also shows the importance of monitoring for countermeasures, as this could have a huge impact on whether deception is being measured accurately. Further, this shows how movement, whether it is deliberate or accidental, can influence the accuracy of the measure. Therefore, this may call into question results from studies where extraneous movement is not accounted for or controlled. Apart from this discussion of countermeasure, limitations of the method are not discussed.

In a study conducted by Kozel et al. (2004), participants were required to go into either one of two rooms, one being a ‘deception’ room and the other being a ‘truth’ room. In both rooms the participants were instructed to find a $50 bill, which was placed under one of the five objects in the room, and remember its location. They were then placed in an fMRI scanner. Those participants in the truth room condition were asked to report the accurate location of the money, while those in the deception room condition were asked to choose an inaccurate location. Location was confirmed using yes/no affirmation to pictures of the objects in the rooms. Participants were told that payment was based on whether their lying could be detected by an ‘investigator’, in order to increase motivation. Results found that there was significant activation in the orbitofrontal cortex and anterior cingulate gyrus. However, there was high variation across the participants. In terms of the type of explanation and conception of the mind, this study follows the same pattern as previously discussed. Deception here is defined as “the purposeful misleading of another” (p. 295). The intent of the deception here is to mislead the investigator about the location of money, and the behaviour includes lying about its location and withholding information about its correct location (though these behaviours were analysed together). There are a large number of limitations identified here. These include the small, male-only sample, the low statistical power of the method, and the liberal threshold for significance. Further, the lesser stakes for the participants (not winning 50 dollars) compared to potential real-life situations was identified as a potential threat to ecological validity.

A further study by Kozel et al. (2005) took a different approach, participants were instructed to complete a mock crime by ‘stealing’ either a watch or a ring in a drawer then placing it in a locker that contained the participant's belongings. They were separately asked whether they had taken the watch and whether they had taken the ring, and again were told that payment was contingent on deceiving the investigator. Results found activation in the anterior cingulate, orbitofrontal cortex and dorsolateral prefrontal cortex. However, similarly to their previous study there was high variation in the activation of these areas, with some subjects failing to show significant activation in some or (in one case) all of these areas. Similar to the studies previously discussed, the conception of the mind is not explicitly considered or stated. The investigation is based on discovering the modular neural correlates of brain activity, meaning that it is a compositional and mechanical explanation. Deception is defined as "a complex process requiring suppression of the truth, communication of a coherent falsehood, contextual knowledge, of that falsehood, and modifications of behaviours to convince the receiver of ones actions." (p. 605). In this study, the intent is to deceive the investigator about whether they took the item, as opposed to the previous study where the intent was to hide where the item was located. The behaviour included lying about their actions, which relies upon episodic memory. This study did determine that participants seemed invested in the outcome of the study, as when asked the majority stated that they believed they were involved in a crime and were motivated by the monetary reward, providing some support for ecological validity. Further, a small number tried to perform countermeasures. However, the authors do identify that this is still far removed from a real-life situation, and the level of risk involved is much lower. Additionally, the authors highlight the fact that the sample used are likely to have very different social and medical histories, and that this experiment requires a cooperative subject.

Langleben et al. (2002, 2005) used a 'Guilty Knowledge Test' (GKT) to investigate the neural correlates of deception. Their specific methodology involved giving the participant a playing card, and asking them to conceal the identity of the card. During imaging, the participant is shown the card and is asked whether they have it. They are also shown other cards, which they have been instructed to tell the truth about (or otherwise forfeit their reward). Results for Langleben et al. (2002) found significant activation in the anterior cingulate gyrus, right superior frontal gyrus, dorsolateral prefrontal cortex, and anterior parietal cortex. Langleben et al. (2005) modified the GKT by adding two stimuli, a recurrent distractor and a null stimulus (back of a card). Results showed activation in the inferior lateral prefrontal and medial superior frontal cortices. Interestingly, no activation in the anterior cingulate cortex was found. Both studies appear to take the default CTM approach, as the theoretical orientation is not discussed. Further, explanation is similarly compositional and mechanistic. Langleben et al. (2002) define deception as “intentional negation of subjective truth” (p. 727). Langleben et al. (2005) define deception as “denying what is” and “an intentional effort to instil a false belief” (p. 262). The experimental protocols are similar in both studies, wherein the intent is to deceive the experimenter about the identity of the card that they hold, and the behaviour is withholding information. Langleben (2002) discusses limitations, including that in this study there is less choice involved in deception (as the participants are essentially only able to deceive in one way) and that there is less guilt and risk. Other limitations include the low statistical power, lack of clarity around the impact of having different shape and colours of stimuli, and use of a single imaging technique (fMRI) rather than in combination with other techniques such as event-related potential or polygraph recordings. Limitations discussed by Langleben (2005) include the low ecological validity and again the lack of choice.

In a study by Spence et al. (2001), participants were instructed to answer questions about whether they had performed certain acts during that same day with either the truth or a lie, depending on whether the question was coded with green or red colouring. All questions were answered with both a truth and a lie. Activation was found in the ventrolateral prefrontal cortex, the medial premotor cortex, the left lateral premotor cortex, and the left inferior parietal cortex. Theoretical orientation and type of explanation are the same as the previous studies discussed. Deception is defined as:

> "Lying is the process by which the perpetrator (of a lie) deliberately attempts to convince their victim of the truth of a proposition they themselves know to be false. A concomitant feature is the liar's withholding of data that they know to be true" (Spence et al., 2001, p. 2849)

The intent here is to deceive the experimenter about whether they had performed an action, utilising episodic memory. The behaviour is simply lying about their actions. The first limitation addressed in this study is that participants were only allowed to make one single motor response when lying, that they did not choose themselves. In a real-life scenario, lying may be more spontaneous. The second limitation was that the questions asked were not emotive, and the task was low-risk and low-stake.

Cui et al. (2013) also used the GKT to investigate deception. The method in this case involved a 'mock murder game'. One participant was designated as the 'murderer' and was required to craft a narrative around a murder and imagine that they were committing it. Six details from this narrative were later used as probes in the GKT. The 'murderer' and the remaining participants ('innocents') were required to memorise a number of target words before undergoing the scanning procedure. In the GKT, participants were asked to indicate whether they had seen a certain word before – this word could be the probes (words associated with crime narrative), the targets (the memorised words), or irrelevant words. The 'murderer' was told to respond negatively to the probes (to conceal their identity) and the 'innocents' were told to respond truthfully. Results showed that the right ventrolateral prefrontal cortex was the most important in categorising murderers and innocents, with

81.25% correct categorisation. Interestingly, the GKT here was modified to provide a ‘judgement’ about whether the participant’s response was truthful or deceptive (the judgement itself was predetermined for the probes and irrelevant words, and unrelated to the response given). During this period of judgement, the researchers found that activity in the right inferior parietal lobe could categorise 93.75% of murderers and 87.5% of innocents correctly, particularly following ‘successful’ deception. Conceptualisation and type of explanation follow the previous trend. No formal definition of deception is included in this study. Deception is instead operationalised as a concealed knowledge task, wherein the participant’s intent is to deceive the experimenter about familiarity with certain words, and the behaviour included withholding information. Further, the results regarding the period of judgement indicate the presence of some potential cognition or emotion following a successful attempt at deception. However, as the researchers indicate, it is not clear whether this is specifically related to deception or some effect of an attentional response following an unexpected stimulus. Aside from this, there were no limitations discussed.

Lastly, a recent study by Ofen et al. (2017) looked at deception using questions designed to tap both episodic knowledge, and personal beliefs and opinions. The participants were given an instruction to either lie or tell the truth and were then presented with a yes/no question. Areas implicated for both types of questions included the bilateral parietal and bilateral medial superior frontal regions. For the episodic questions only, the right temporal pole was differentially activated, and for the belief questions the precuneus was relevant. This study does not deviate from the pattern set by the previous in terms of its lack of discussion about the theoretical orientation of the research and the compositional and mechanistic nature of explanation. The definition used here is extensive:

> "The definition of the verb 'lie' according to the Merriam-Webster dictionary is 'to make an untrue statement with intent to deceive'. Implicit in this definition are three facets of the act of lying. First, an individual must know or determine the truth in order to lie about it. Depending on the nature of the lie, this step could involve remembering specific details of an event or past experiences, or assessing one's opinion or emotional state. Second, a deceiver must then prepare an untrue statement, and finally, respond by providing the statement that is contrary to his or her determined truth. On a cognitive level, these last two steps may involve cognitive control so as to inhibit the truthful response or manipulate the true information to convert it into a lie" (Ofen, et al., 2017, p. 117).

The participant's intent in this study is to lie about things that they have done or things that they believe. Alternatively, it could be argued that a concurrent intent here is to follow the instructions, rather than actually try to deceive another person. Whereas in other studies attempts have been made to raise the stakes (usually using a monetary reward) so that participants will be more encouraged to deceive the experimenter, it is not clear in this study that there is equivalent motivation. Therefore, the behaviour of lying is present, but the actual intention is not obvious. It is important to be clear about this, because it is not possible to separate out these two intents post-hoc, and it could have intrinsic effects on the validity of the data. A number of limitations are highlighted by the authors. The first is that being asked to lie or tell the truth is essentially a task where the participant is trying to obey the experimenter, rather than in a real-life situation where the participant would be actively trying to subvert the experimenter (lacking ecological validity). Second, it is possible that measures taken to exclude the emotional component of deception were not completely successful, given that regions associated with emotion were activated during this task. However, in a real-life situation emotion could be an integral part of the deception process. Thus it is not clear whether this would be a limitation because emotion was included, or a limitation because the emotion felt by the participants was not as extensive as it could be in real-life; therefore downplaying the role of these emotion-processing areas. A third limitation specified by the authors is that the questions about past experiences were limited by the potential strength of the memory.

Conclusions

Table 1 provides a summary of the literature concerning deception, with a tick symbol representing the idea that the study addressed it well, and a cross representing if a study addressed it poorly. The following is clear overall: 1) there is little consideration for the theoretical orientation that underlies the research; 2) that the concept of deception is not well defined; 3) that the experimental paradigms are not consistent over different studies and research labs (or even within the same group of researchers across studies); and 4) there is an overall inconsistency and huge amount of variability of brain region activation, likely existent due to the previous points.

Table 1. Summary of deception literature

	Consideration			
	Theoretical	Explanation	Definitional	Methodological
Davatzikos et al. (2005)	×	×	×	✓
Volz et al. (2015)	×	×	✓	×
Ganis et al. (2003)	×	×	✓	✓
Ganis et al. (2011)	×	×	×	×
Kozel et al. (2004)	×	×	✓	✓
Kozel et al. (2005)	×	×	✓	✓
Langleben (2002)	×	×	✓	✓
Langleben (2005)	×	×	✓	✓
Spence et al. (2001)	×	×	✓	✓
Cui et al. (2013)	×	×	×	×
Ofen et al. (2017)	×	×	✓	✓

How is the mind conceptualised?

In terms of theoretical orientation, it is clear that there is little to no emphasis by researchers on explaining the theory that underpins their studies. It is assumed by the content of the research that the 'default' modular CTM approach is being used, as there is a large focus on detecting isolated regions of activity. There is occasional reference to the importance of neural networks (e.g., Volz et al., 2015).

What type of explanation is being used?

The type of explanation used in these studies is consistently compositional and mechanistic. This is not a problem in itself, however it should be made explicitly clear what the researchers are trying to explain and how they aim to explain it. This may ward against others using this research to make unfounded causal claims. In terms of the levels of explanation used in this research, the neuroscientific level was heavily emphasised.

How is the concept of deception defined and operationalised?

The definitions used in the literature are not consistent at all. In fact, no single definition was used in two pieces of research, even if the research was by the same researchers. Some research included no definition. Some definitions were more extensive, breaking down the concept of deception into its different components (Ofen et al., 2017).

Similarly, there was a lack of consistency in the experimental paradigms used between researchers. There were some similarities – for example, a number of studies used a variation of the GKT. However, there does not seem to be a single 'gold standard' experiment for investigating deception, or an attempt to standardise how deception is operationalised. Therefore, it is difficult to compare the results gained by each study as even if the definitions were similar, the way in which the experimental protocols were designed could mean that the type of deception or the actions performed by participants in the deception process were different.

What are the methodological and conceptual issues evident and how are they accounted for?

There is some discussion of the limitations of deception detection in the literature. There is a large focus on the technical limitations. These include the following: that it is time consuming to collect the baseline data (Volz et al., 2015); that irrelevant brain areas (noise) may be implicated (Cui et al., 2013, Volz et al., 2015); that these results are based on group analysis rather than a single participant (Ganis et al., 2003); that countermeasures may be used (Ganis et al., 2011; Kozel et al., 2005); that the samples are small and gender biased (Kozel et al., 2004); that the participants will likely have different social and medical histories (Kozel et al., 2005); that many of the studies use a forced-choice paradigm (Langleben et al., 2002; Spence, 2001); and that the strength of memory may be different for participants recalling an unimportant event, as opposed to a real-life situation (Ofen, 2017).

There are also important conceptual limitations discussed. Several of these studies discuss the issue of the potential impact of emotions such as fear and guilt on the validity of the measure (Ganis et al., 2003; Langleben et al., 2003; Ofen et al., 2017; Spence et al., 2001). Ganis et al. (2003) claims that neural correlates of deception can solve the issue of traditional polygraph lie detection being seen as ‘guilt detection’ because of the idea that increased arousal (which is what is measured by the polygraph) can be interpreted as emotions such as guilt and fear rather than an indication of deception. However, the studies by Ganis et al. (2003) and others do not account for the fact that these emotions are likely to be present during deception, and this will be represented in the neural correlates. Whether these emotions were present in these studies is not possible to establish, but it is likely due to the lack of emotionally probing content that emotions were at least dulled, if not absent. This is particularly important if applied to a situation where a defendant's future could be partially determined by the outcome. This relates to another conceptual issue discussed in the literature, concerning whether the experiment accurately captures the risk involved in a real-life situation (Kozel et al., 2004; Langleben et al., 2002; Spence et al., 2001). It is unlikely that a laboratory experiment could provide an accurate imitation of real-life risk and the severity of the possible outcomes of deception. Lastly, Ofen et al. (2017) raised the issue that many of these experiments could be at least partially measuring obedience to the experimenter. On the other hand, in a real-life situation the participant would be actively trying to deceive for their own benefit (e.g., to avoid incarceration). Therefore, this calls into question the ecological validity of the experiment.

In summary, areas implicated included: right prefrontal cortex, bilateral posterior cortex, right temporoparietal junction, superior temporal gyrus, precuneus, bilateral cuneus, dorsal anterior cingulate cortex, middle frontal gyrus, fusiform gyrus, left cerebellum, bilateral precentral gyrus, insula, inferior parietal lobe, thalamus, caudate nucleus, lenticular nucleus, amygdala, hippocampus, dorsolateral prefrontal cortex, orbitofrontal cortex, anterior parietal cortex, ventrolateral prefrontal cortex, and the right temporal pole. Therefore, according to these studies virtually the whole brain is implicated in deception research. Figure 1 shows a diagram by Uttal (2011) that illustrates this point succinctly. Additionally, while there is some overlap in the regions activated, there is wide variety in terms of which of these areas are activated during each experiment. There seems to be some indication that the anterior cingulate cortex, and various prefrontal/frontal/parietal areas are important, but again this is not consistent – one study found no activation in the anterior cingulate cortex (Langleben, 2005). It is clear that there are huge problems in the research on deception detection and overall, it is clear that the area is not currently robust enough for use in forensic psychology.

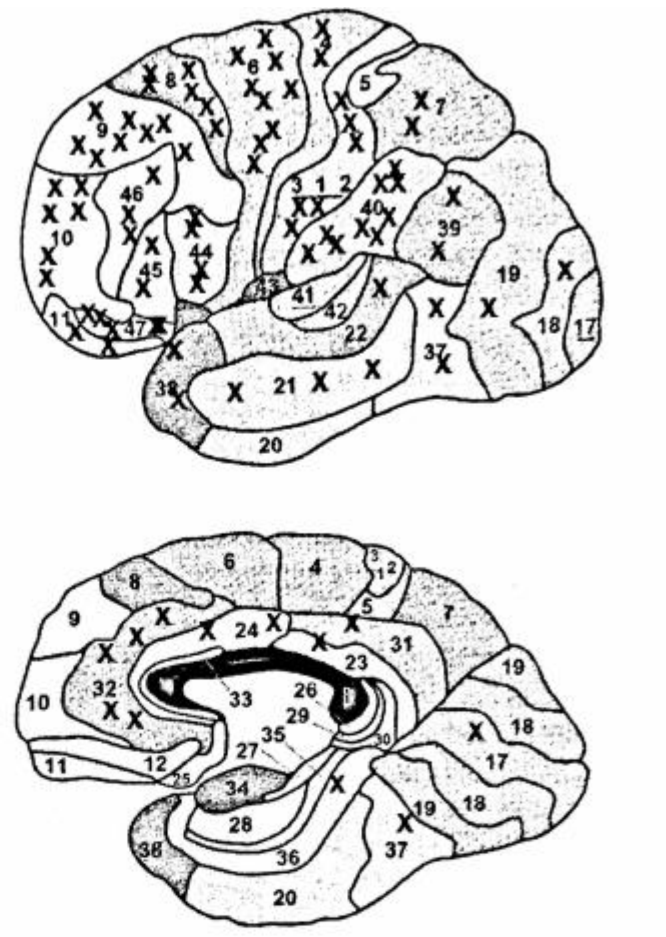

Figure 1. Distribution of activations during deception. Reprinted from *Mind and brain: A critical appraisal of cognitive neuroscience* (p. 338) by W. R. Uttal, 2011, Cambridge, MA: MIT Press. Copyright (2011) by William Uttal.

CHAPTER 4: Mind Reading by Brain Reading Case Studies

The concept of 'mind reading' refers to the idea that we may be able to tell what a person is thinking by scanning their brain activity and inferring meaning from the activity patterns. In the forensic area, this could be considered useful for determining knowledge about a crime. As with deception, I will analyse the literature based on the following questions: 1) how is the mind conceptualised; 2) what type of explanation is being used; 3) how the concept of mind reading is defined and operationalised; and 4) the methodological and conceptual issues evident and how are they accounted for.

Mind reading in the neuroscience literature generally focuses on perception-based experimental protocols, largely using activity gleaned from the visual areas in the occipital lobe of the brain. Therefore, in terms of assessing the consistency of definitions and operationalisations between laboratories it is important to consider the following: 1) the task itself and the behaviour of the participants; 2) the stimuli used; 3) the neuroscientific method used to determine brain activity 4) the outcome of the measurement.

Literature Analysis

In the interest of conciseness, some similarities of the below studies have been summarised. Unless otherwise stated, in terms of the conceptualisation of the mind, these studies follow the pattern of not clarifying the theoretical background, and therefore it can be assumed that they follow CTM. Further, these studies use compositional and mechanistic explanation, based on a single level of explanation – largely the neural activity in the occipital cortex.

A study by Kamitani and Tong (2005) found that by looking at activity patterns in visual areas, they could accurately predict which of several differently oriented gratings was being viewed at the time. Participants in the study were required to view one of eight stimulus orientations for each 16 second trial. The experimenters used this information to construct an 'orientation decoder', which was then tested using samples that were independent from those used to train the decoder. Following on from this, the experimenters were interested in whether this could apply to the area of mind reading. Participants were given two stimuli with different orientations which were used to train the decoder. Then participants viewed a composite image of both stimuli and were asked to attend to one or the other. Activity patterns accurately predicted which stimuli orientation was being attended to with 80% accuracy. In this study, mind reading is defined as whether the stimulus orientation a person is viewing can be revealed "determining a subject's mental state given knowledge of his or her brain state" (p. 682). This is a visual attention task, where the stimuli are two differently oriented images and the participants are required to attend to one or the other. The method used is a decoder that has been trained to detect the pattern of brain activity associated with attending to certain stimulus orientations. The outcome of the task is whether the decoder can predict which stimulus is being attended to. The limitations of this method are not discussed. In fact, the authors posit the following:

"Our approach may be extended to studying the neural basis of many types of mental content, including a person's awareness, attentional focus, memory, motor intention and volitional choice." (Kamitani & Tong, 2005, p. 684).

The idea that this research on the correlates of visual subjective experience can generalise to a wide range of perceptual and cognitive processes is a huge leap.

Similarly, Seymour et al. (2009) found that they could discriminate between different combinations of colour and motion using similar visual areas. In this study, participants were required to view a display with red or green coloured dots rotating around the centre. Using a pattern analysis, the experimenters were able to determine the pairing of colour (red/green) and motion direction (clockwise/counter-clockwise) using the activity patterns in the visual cortex with accuracy ranging from 57% to 86% and 57% to 93% respectively. Mind reading is defined as the specific pairing of colour and motion direction decoded from BOLD responses. The task itself is passive, with the participants required to view different colours and motion directions. This experiment used a multivariate pattern analysis that did not require previous training. The outcome related to predicting the pairing of the stimuli. The authors do acknowledge the limitations of fMRI, in that the specific neural mechanisms underlying activation cannot be discovered.

A study by Haynes, Deichmann and Rees (2005) found that it was possible to determine which stimulus was dominant during binocular rivalry. Regions within the lateral geniculate nucleus were shown to exhibit a preference for either right or left eye stimulation. Mind reading was defined as “whether the human LGN shows eye-specific changes in signal in association with behaviourally measured changes in perception during binocular rivalry” (p. 1). This was a binocular rivalry task, which used two different gratings (one vertical and one horizontal). Participants attended to either the left or right eye stimuli and indicated which one they were focused on using a button press. Activity was analysed using a General Linear Model and voxel analysis. The outcome the researchers were interested in was whether attention to either eye stimulus could be reliably predicted. Limitations discussed are restricted to the difficulty of measuring relative suppression of images during binocular rivalry.

A study by Polyn et al. (2005) found that they could discriminate between the categories of faces, places, and common objects that the participants were thinking of during free-recall. This was based on the contextual reinstatement hypothesis: the idea that the pattern of brain activity during recall of an episodic memory will resemble the pattern of activity during the episode itself. While in the fMRI machine, participants were required to study 30 items (each item containing a photograph and a name) of the aforementioned categories. During the free recall period the participants recalled all of these items, in any order. Results showed that in seven of the nine participants a significant effect was found: that the classifier correctly categorised the recall (for example, if the person was thinking of a face, the classifier matched this brain activity to the ‘face’ category). In contrast to the other studies reviewed here, this study does explain the theory behind their research, in that it is based on neural network theory and the contextual reinstatement hypothesis. However, the type of explanation here is still in step with the other research. In this study, mind reading is explained as “when subjects try to recall specific details from a particular episode or type of episode, the pattern of brain activity (during recall) will progressively come to resemble the pattern of activity that was present during the to-be remembered episode” (p. 1963). The activity is both visual and memory based, with three different categories of visual stimuli. Categorisation was analysed using a neural network classifier trained for each individual to discriminate between the categories. The intended outcome is to be able to accurately distinguish between these categories. Technical and conceptual limitations are not covered here. In fact, the authors refer to their results as “neural fact”. This does not account for the multitude of potential confounding factors, particularly considering that fMRI is a secondary measure of response.

Remarkably, there have also been studies that have shown the ability to reconstruct images of what people are thinking. Miyawaki et al. (2008) used a task where participants were shown a 10 by10 grid with flickering checks that formed simple shapes and letters. By analysing brain activity, they could reconstruct which shape or letter the participant was viewing. Interestingly, while the reconstructed images had similar essential features to the shapes used to train the decoder, the images given to participants and subsequently reconstructed were unfamiliar to the decoder. Mind reading in this experiment was operationalised as the brain activity present when viewing a visual stimulus. This was a passive experiment, as the participants were only required to view the image and indicate what they saw. The stimuli were simple, black and white shapes and letters. Activity was analysed using a decoder trained on participants using a large number of shapes to associate certain voxel activity patterns with stimuli features. The outcome here was reconstruction of the image stimuli. The authors discuss some of the limitations of their method, including that signals could be spontaneously induced, particularly for neighbouring voxels, movement and physiological noise (such as cardiac or respiratory), and the condition of the fMRI scanner.

Similarly, Nishimoto et al. (2011) showed that a movie stimulus could be reconstructed from the BOLD signal. The reconstruction is imperfect due to issues with noise and the spatial and temporal resolution of the technique, but the results are quite striking, in that the pictorial reconstructions clearly align with the original stimulus. The task here is passive viewing of a movie stimulus. The stimuli were natural movie clips, and similar clips were used to train the model. The outcome was that the model was able to reconstruct the clips seen, as long as the clips were similar to those used to train the model. The authors do discuss the limitation that reconstructions were poor if the viewed clip did not contain scenes that were similar to the sample clips. Further, they acknowledge the issues with spatial and temporal resolution for fMRI and that this impacted on the quality of the reconstruction.

There have also been studies that have shown the ability to predict intention. For example, Haynes et al. (2007) found that they could predict, with roughly 70 percent accuracy, whether a participant was intending to perform an addition or subtraction task. The participants in this case were able to choose which task they wanted to perform. Brain responses were measured during this period of selection and were analysed using a pattern classification method. Mind reading here was defined as “whether the current intentions of a subject were encoded in specific regions of prefrontal cortex” (p. 1). This was restricted by the researchers to the prefrontal cortex. Though no reason was given for this restriction, it can be assumed that this was because previous research has indicated the importance of the prefrontal cortex in intentionality. However, this is problematic as it excludes a large portion of the brain a priori and restricts the validity of the results. Unfortunately, the authors do not discuss this limitation or explain why their research is conceptually valid despite this restriction. A limitation pointed out by the authors was that the rapid pacing of trials made it difficult to differentiate between the period before task selection and the task selection itself.

Conclusions

The previous studies have shown the following overall: 1) again we see that the theoretical background of the research is not explicitly acknowledged; 2) that the concept of mind-reading is defined variably amongst researchers; 3) there are a wide array of experimental paradigms; and 4) conceptual limitations of the various techniques are not considered, but there is some consideration of technical limitations.

Table 2 summarises how well each study covered the four areas of consideration, with a tick symbol representing the idea that the study addressed it well, and a cross representing if a study addressed it poorly.

Table 2. Summary of mind-reading literature

	Consideration			
	Theoretical	Explanation	Definitional	Methodological
Kamitani and Tong (2005)	×	×	✓	×
Seymour et al. (2009)	×	×	×	✓
Haynes et al. (2005)	×	×	✓	✓
Polyn et al. (2005)	×	×	✓	×
Miyawaki et al. (2008)	×	×	×	✓
Nishimoto et al. (2011)	×	×	×	✓
Haynes et al. (2007)	×	×	✓	×

How is the mind conceptualised?

As summarised previously, there is little attention paid to the theoretical background of the research in this area, as was found with the deception research. The only exception to this seemed to be Polyn et al. (2005), though this was not extensive. The literature appears to implicitly follow a modular, CTM approach.

What type of explanation is being used?

The research was wholly compositional and mechanistic, and focused heavily on the neuroscientific level. This means that causal claims cannot be made from this research, and that there is a lack of data provided for other levels of analysis. However, due to the nature of mind-reading (as it attempts to capture a perceptual process at a certain moment in time), this is to be expected.

How is the concept of mind reading defined and operationalised?

From these studies, it is clear that there is wide variation in the way the concept of mind reading is researched. Even for the studies that focused on the visual areas of the brain, the tasks themselves were varied, and included visual attention, binocular rivalry, a memory task, passive viewing of motion/colour, shapes/letters, and a movie stimulus. This also impacted on the definition of mind-reading used, as it was taken to be whatever brain activity that was analogous to the chosen task. This provides a problem for the development of any mind-reading technique, as these studies are not directly comparable, and therefore any technique developed would not have a strong base in the literature.

What are the methodological and conceptual issues evident and how are they accounted for?

The range of limitations given by the literature is not extensive. Two of the above studies do not cover the limitations of their method (Kamitani & Tong, 2005; Polyn et al., 2005).

Seymour et al. (2009) and Haynes et al. (2005) include the limitations of fMRI as a secondary measure rather than one that directly measures neural activity. The other three consider the problems of noise and low spatial/temporal resolution of fMRI (Haynes et al., 2007; Miyawaki et al., 2008; Nishimoto et al., 2011).

There are a number of other practical and conceptual limitations to the techniques that need to be addressed. The first issue is the reliance on the visual areas of the brain – most of the research into brain-reading techniques use a visual stimulus and observes the occipital lobe; which has many well-defined areas (Tong & Pratte, 2012). For example, early and late visual areas can be differentiated, allowing researchers to easily discriminate between specific visual activities (Hochstein & Ahissar, 2002). Additionally, the act of viewing an image is much simpler to operationalise than other cognitive tasks, whereas for beliefs or social interactions it may not be as easy to create a task and assess brain activity. This is a particularly important issue in neuroscience, due to the limitations of functional imaging. While in an fMRI, the subject must lay completely still. Tasks usually involve viewing a 2D image or listening to an auditory stimulus. This is a problem for the ecological validity of the concept that is being assessed, particularly as the concept becomes more complex (Sullivan, 2015). Therefore, while the evidence for the ability to 'read' what a person is viewing is quite strong, there is little evidence that this can practically be extrapolated to more complex concepts such as mental states.

Another issue is the reliance on knowing the original stimuli. For example, the findings by Haynes et al. (2007) rely on the fact that they knew that the person would either be thinking of addition or subtraction. There is no way to infer their intention to add or subtract without this prior knowledge. Furthermore, their findings also rely on accurate confirmation by the participant that their prediction was correct. Again, reverse inference is not possible. This may be particularly relevant in criminal cases where the stakes are higher, and therefore there is more incentive for the defendant to provide incorrect information. Further, the studies are quite resource intensive, as the majority require the training of a decoding device in order to accurately classify the brain signals.

While these issues are important, they are technical limitations that could potentially be overcome by technological developments. Therefore, it is also important to understand what brain-reading cannot do conceptually. The brain can be thought of as the 'realiser' of the mind, meaning that we can see the areas that are physiologically analogous to psychological phenomena (Yaffe, 2016). However, this is also a limitation of the technique, as it does not allow us to examine past nor future brain activity and thoughts. This is particularly relevant for criminal trials, as the intention is often to determine what a person was thinking during the crime itself, which may have taken place months or even years earlier. Whilst memory could be used to indicate past events, this information can become lost or distorted over time, and there is no guarantee that the memory is accurate (Yaffe, 2016). Studies have shown that human memory is hugely malleable. For example, a study by Loftus and Pickrell (1995) showed that a false memory could be implanted using just the suggestion of a plausible story, to the point where the person believed the memory to be true even when told that it had been implanted. Considering that a criminal defendant may not have a perfect memory of the crime in the first case (particularly if drugs or alcohol were involved), and may have been influenced by other information, the use of brain-reading techniques could provide unjustified authorisation of these memories and should therefore be cautioned. Furthermore, because mental states and cognitions change so rapidly, and are influenced by factors outside of the person (such as an interaction with another person), there is no way to determine what state the brain will be in even the next day, let alone any later date in the future (Yaffe, 2016). Therefore, even if technological developments allowed a further refinement of mind-reading techniques, it is conceptually impossible to be able to 'read' the minds of criminal defendants in a way that is practically useful for determining criminal responsibility. Lastly, the absence of activation cannot be taken as proof a deficit or inexistence of a thought or behaviour (Brown & Murphy, 2010). Therefore, in terms of criminal responsibility, brain-reading techniques could not be used to prove innocence, only guilt, reducing the overall applicability**.**

Besides the issues already explored in both the deception and brain-reading literature, it is also important to keep in mind that these studies have been performed on the general population. Considering that the forensic population has been shown to have high rates of mental disorders such as antisocial personality disorder and depression (see Fazel & Danesh, 2002) and potentially have fundamental structural differences in their brains, it is highly likely that the results would not easily generalise to the target population. For example, a meta-analysis conducted by Yang and Raine (2009) found that there were significant prefrontal structural and functional impairments in antisocial individuals. Additionally, a study by Barkataki et al. (2006) found that individuals with antisocial personality disorder exhibited reduced whole brain volume and reductions in the temporal lobe and putamen. As Yaffe (2016) explains, the mapping of the statistical information to the canonical brain is already imprecise, and therefore any lesions or differences in brain structure or function could decrease the precision even further.

Therefore, in order to introduce techniques such as those discussed above, the following would need to occur: 1) the construct would need to be accurately and consistently defined across the discipline; 2) experimental methodology would need to be standardised; 3) these standardised techniques would need to be repeated on a diverse offender population; 4) the accuracy and reliability would need to be sufficiently high to avoid false positives.

CHAPTER 5: Discussion and Conclusions

The area of forensic neuroscience is becoming increasingly prevalent, and it is therefore important to ensure that research is based on a strong and valid theoretical foundation. It is particularly important given the practical applications of forensic psychology, and the very real effects that research can have on people’s lives – for example, effects on sentencing or treatment. The second chapter of this thesis outlined some of the conceptual, theoretical, and practical issues that researchers should acknowledge when integrating neuroscientific research into the forensic area. These include: considerations about theory and conception of the mind, explanation, concept validity and validity of experimental protocols, and technical limitations. The third and fourth chapters have investigated whether these considerations are reflected in the subject areas of deception detection and mind-reading by brain-reading. This chapter will explore some of the potential strategies that researchers could employ in order to mitigate these issues.

Solutions

Theory

The first issue that was outlined was the lack of emphasis placed on the theoretical background of neuroscientific research. Much research is based on reverse inference from neural activity to concepts, rather than being directed by theory. As Ward et al. (2018) explain, this is important because theory is central to every stage of scientific inquiry – theory affects how hypotheses are constructed and selected, which participants are used, how concepts are defined, how experimental protocols are designed, and how results are interpreted. The research in the areas of deception and mind-reading showed that there is currently little consideration for the theoretical background that underpins the research. A solution here is to make sure that research is based on a strong and explicit theoretical framework. One example of this, the Dynamic Risk Research Framework (DRRF), was developed by Ward and Fortune (2016) in order to guide investigation of Dynamic Risk Factors. Currently in forensic psychology, the most common framework for explaining and treating offender behaviour is the RNR model and its reliance on dynamic risk factors (DRFs) as causal phenomena. However, several researchers in recent years have highlighted a number of issues with using DRFs as explanations for offending, including the problem that they have vague definitions, conflate explanation and prediction, and may have low construct validity (Ward & Beech, 2015; Ward & Fortune, 2016). The DRRF is an effort to move the field from a reliance upon descriptions of predictors to a causal explanation consistent with existing theory and empirical evidence. Table 3 provides an overview of the framework. Each risk factor is broken down via two criteria: the causal processes underlying the risk factor, and the different levels of analysis. This framework therefore encourages both breadth and depth of explanation, leading to a stronger and more complex explanation overall.

Table 3. Dynamic risk research framework.

Dynamic risk factor	Intimacy Deficits (e.g. emotional congruence with children)			
	Levels of analysis			
Casual processes	Biological	Behavioural	Phenomenological	Contextual
Negative affective systems				
Cognitive systems				
Intrapersonal social processes				
Self-regulation				
Interpersonal social systems				

Note: Adapted from “From dynamic risk factors to causal processes: a methodological framework,” by T Ward and C Fortune, 2016, *Psychology, Crime & Law*, *22*, 190-202.

Explanation

The second issue discussed involves the kinds of explanation used in research. Explanatory pluralism is important in order to create explanation of phenomena that are comprehensive and theoretically valid (Manzavinos, 2016). It is not clear whether researchers are aware of the differences between different kinds of explanation, as evidenced by the deception and mind-reading literature, where there was no indication of the type of explanation that was occurring. Additionally, all of the research was compositional, as opposed to etiological. A potential solution for this is to explicitly differentiate between compositional and causal explanations. For example, if the research is compositional, the researchers should explain that their results should not be used to make causal claims. Additionally, if a piece of research is making causal claims, this should be underpinned by strong theory. Further, research should endeavour to employ explanatory pluralism when possible.

Similarly, there was an issue with research overemphasising the neurobiological level of explanation, and underemphasising other equally important levels such as the phenomenological, behavioural, and social. This was clear in the deception and mind-reading areas, as they largely focused on the neurological level. To avoid reductionist leanings, neuroscientific experimenters should acknowledge that their research is a part of the explanatory mosaic and attempt to include other levels when discussing the importance of their findings.

Concept Validity

Third, there was also an issue with concept validity. Definitions are vague and inconsistent between different studies. This is a particular problem in the deception literature, as there were several different definitions used, or even no definition provided. There are a number of potential solutions here. Firstly, researchers should make sure that they are using consistent and clear definitions of concepts within their own lab. Secondly, collaboration with researchers in different labs who are studying the same phenomena could assuage the issues with validity. As posited by Sullivan (2016) and Poldrack (2010) the best solution could be to create a discipline-wide cognitive ontology, wherein concepts and constructs are explicitly defined. For the mind-reading literature, it is more difficult to have a consistent definition as it is a more heterogeneous topic. However, researchers should still strive to be as consistent as possible with prior research or provide explicit definitions for how they define mind-reading.

Methodology

Along similar lines, a fourth issue is the variability in research methodology within psychology. Once consistent definitions of concepts are established, the next step would be to ensure that the experimental protocols being used to explore these concepts are valid (in that each experiment is actually eliciting the phenomena to be studied). As with the problem of inconsistent definitions (and potentially partially because of the wide variety of definitions) in both the deception and mind-reading literature there were huge inconsistencies in the methods used. Researchers within each subfield should strive towards having a consistent methodology. This would allow for simpler comparisons between studies, and increased validity.

Technical Limitations

Furthermore, a fifth issue is the large number of technical limitations within neuroscience. Many of these are essential features of the technology and cannot be changed, except with developments of new and improved technologies. Additionally, these limitations are often considered common knowledge and are therefore not included in discussion of results. Some technical limitations were variously discussed by researchers in the deception and mind-reading literature, including issues with noise, secondary measurements, spatial and temporal resolution, and how resource intensive the methods are. However, when considering bringing this research into the forensic area it is important to be more explicit. This is particularly important if the research is being used in a more practical context, such as in a courtroom, due to the numerous ethical concerns. It is also essential for those parties involved in transferring and translating this research to be aware of the technical limitations and clarify these to naïve audiences. Discussion of the conceptual limitations of the methods were much less common. This is surprising considering the large number of conceptual problems with the measurement of brain activity. This could be mitigated by increasing the ecological validity of the studies, so that they more closely resemble a real-life situation. In addition, using a forensic population to test these measures would greatly increase validity.

Conclusion

Overall, there is definite and increasing interest in incorporating neuroscientific research into the field of forensic psychology, and this research could provide valuable insight into human behaviour. Further, it could prove to be advantageous in more practical forensic areas, such as the court system or therapeutic interventions. However, this thesis has shown that we should be wary about integrating this research into a forensic setting in its current state, and that the research needs to be more rigorous and consistent about its theoretical underpinnings, definitions and operationalisation of concepts, and technical and conceptual limitations. Ideally, future studies should take into consideration the issues outlined in this thesis and strive towards having more theoretically valid research. Through this, forensic psychology will continue to flourish as a discipline.

References

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. MIT Press.

Barkataki, I., Kumari, V., Das, M., Taylor, P., & Sharma, T. (2006). Volumetric structural brain abnormalities in men with schizophrenia or antisocial personality disorder. *Behavioural brain research*, *169*(2), 239-247. https://doi.org/10.1016/j.bbr.2006.01.009

Barratt, N (2016). *Brain scanning: Technology to read minds*. Retrieved from http://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=11665719

Brown, T., & Murphy, E. (2010). Through a scanner darkly: functional neuroimaging as evidence of a criminal defendant's past mental states. *Stanford law review*, 1119-1208. https://doi.org/10.1016/j.bbr.2006.01.009

Cheng, Y., Chen, C., Lin, C. P., Chou, K. H., & Decety, J. (2010). Love hurts: an fMRI study. *Neuroimage*, *51*, 923-929. https://doi.org/10.1016/j.neuroimage.2010.02.047

Cui, Q., Vanman, E. J., Wei, D., Yang, W., Jia, L., & Zhang, Q. (2013). Detection of deception based on fMRI activation patterns underlying the production of a deceptive response and receiving feedback about the success of the deception after a mock murder crime. *Social cognitive and affective neuroscience*, *9*(10), 1472-1480. https://doi.org/10.1093/scan/nst134

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D. G., Acharyya, M., Loughead, J. W., ... & Langleben, D. D. (2005). Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage*, *28*(3), 663-668. https://doi.org/10.1016/j.neuroimage.2005.08.009

Deception. (2018). In *Merriam-Webster.com*. Retrieved from https://www.merriam-webster.com/dictionary/deception

Drayson, Z. (2009). Embodied cognitive science and its implications for psychopathology. *Philosophy, Psychiatry, & Psychology*, *16*(4), 329-340. https://doi.org/10.1353/ppp.0.0261

Fisher, H., Aron, A., & Brown, L. L. (2005). Romantic love: an fMRI study of a neural mechanism for mate choice. *Journal of Comparative Neurology*, *493*, 58-62. https://doi.org/10.1002/cne.20772

Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L., & Yurgelun-Todd, D. A. (2003). Neural correlates of different types of deception: an fMRI investigation. *Cerebral cortex*, *13*(8), 830-836. https://doi.org/10.1093/cercor/13.8.830

Ganis, G., Rosenfeld, J. P., Meixner, J., Kievit, R. A., & Schendan, H. E. (2011). Lying in the scanner: covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *Neuroimage*, *55*(1), 312-319. https://doi.org/10.1016/j.neuroimage.2010.11.025

Haynes, J. D., Deichmann, R., & Rees, G. (2005). Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. *Nature*, *438*(7067), 496-499. https://doi.org/10.1038/nature04169

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. https://doi.org/10.1016/j.cub.2006.11.072

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*(5), 791-804. https://doi.org/10.1016/s0896-6273(02)01091-7

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., ... & Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679-685. https://doi.org/10.1038/nn1444

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *361*(1476), 2109-2128. https://doi.org/10.1098/rstb.2006.1934

Kolber, A. J. (2016). Free will as a matter of law. In Patterson, D., & Pardo, M. S. (Eds.), *Philosophical Foundations of Law and Neuroscience*. (pp. 9-28) Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198743095.003.0002

Kozel, F. A., Johnson, K. A., Mu, Q., Grenesko, E. L., Laken, S. J., & George, M. S. (2005). Detecting deception using functional magnetic resonance imaging. *Biological psychiatry*, *58*(8), 605-613. https://doi.org/10.1016/j.biopsych.2005.07.040

Kozel, F. A., Revell, L. J., Lorberbaum, J. P., Shastri, A., Elhai, J. D., Horner, M. D., ... & George, M. S. (2004). A pilot study of functional magnetic resonance imaging brain correlates of deception in healthy young men. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *16*(3), 295-305. https://doi.org/10.1176/appi.neuropsych.16.3.295

Langleben, D. D., Schroeder, L., Maldjian, J. A., Gur, R. C., McDonald, S., Ragland, J. D., ... & Childress, A. R. (2002). Brain activity during simulated deception: an event-related functional magnetic resonance study. *Neuroimage*, *15*(3), 727-732. https://doi.org/10.1006/nimg.2001.1003

Langleben, D. D., Loughead, J. W., Bilker, W. B., Ruparel, K., Childress, A. R., Busch, S. I., & Gur, R. C. (2005). Telling truth from lie in individual subjects with fast event-related fMRI. *Human brain mapping*, *26*(4), 262-272. https://doi.org/10.1002/hbm.20191

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric annals*, *25*(12), 720-725. https://doi.org/10.3928/0048-5713-19951201-07

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition*, *107*(1), 343-352. https://doi.org/10.1016/j.cognition.2007.07.017

McGugin, R. W., Newton, A. T., Gore, J. C., & Gauthier, I. (2014). Robust expertise effects in right FFA. *Neuropsychologia*, *63*, 135-144. https://doi.org/10.1016/j.neuropsychologia.2014.08.029

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M. A., Morito, Y., Tanabe, H. C. & Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, *60*(5), 915-929. https://doi.org/10.1016/j.neuron.2008.11.004

Morse, S. J. The inevitable mind in the age of neuroscience. In Patterson, D., & Pardo, M. S. (Eds.), *Philosophical Foundations of Law and Neuroscience*. (pp. 29-50) Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198743095.003.0003

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641-1646. https://doi.org/10.1016/j.cub.2011.08.031

Ofen, N., Whitfield-Gabrieli, S., Chai, X. J., Schwarzlose, R. F., & Gabrieli, J. D. (2017). Neural correlates of deception: lying about past events and personal beliefs. *Social cognitive and affective neuroscience*, *12*(1), 116-127. https://doi.org/10.1093/scan/nsw151

Pardo, M. S., & Patterson, D. M. (2013). *Minds, brains, and law: The conceptual foundations of law and neuroscience*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199812134.001.0001

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59-63. https://doi.org/10.1016/j.tics.2005.12.004

Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., ... & Bilder, R. M. (2011). The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00017

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310* (5756), 1963-1966. https://doi.org/10.1126/science.1117645

Seymour, K., Clifford, C. W., Logothetis, N. K., & Bartels, A. (2009). The coding of color, motion, and their conjunction in the human visual cortex. *Current Biology*, *19*(3), 177-183. https://doi.org/10.1016/j.cub.2008.12.050

Spence, S. A., Farrow, T. F., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. (2001). Behavioural and functional anatomical correlates of deception in humans. *Neuroreport*, *12*(13), 2849-2853. https://doi.org/10.1097/00001756-200109170-00019

Sullivan, J. A. (2009). The multiplicity of experimental protocols: a challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese*, *167*(3), 511-539. https://doi.org/10.1007/s11229-008-9389-4

Sullivan, J. A. (2010). Reconsidering ‘spatial memory’ and the Morris water maze. *Synthese*, *177*(2), 261-283. https://doi.org/10.1007/s11229-010-9849-5

Sullivan, J. A. (2015). Experimentation in cognitive neuroscience and cognitive neurobiology. *Handbook of Neuroethics*, 31-47. https://doi.org/10.1007/978-94-007-4707-4_108

Sullivan, J. A. (2016). Construct stabilization and the unity of the mind-brain sciences. *Philosophy of Science*, *83*(5), 662-673. https://doi.org/10.1086/687853

Thagard, P. (2016). *Brain-mind: From neurons to consciousness and creativity*, draft 4, University of Waterloo.

Thagard, P. (2017). *Natural Philosophy: From Social Brains to Knowledge, Reality, Morality, and Beauty*, draft 3, University of Waterloo.

Tong, F., & Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual review of psychology*, *63*, 483-509. https://doi.org/10.1146/annurev-psych-120710-100412

Uttal, W. R. (2011). *Mind and brain: A critical appraisal of cognitive neuroscience*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262015967.003.0001

Volz, K. G., Vogeley, K., Tittgemeyer, M., Von Cramon, D. Y., & Sutter, M. (2015). The neural basis of deception in strategic interactions. *Frontiers in behavioral neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00027

Ward, T., & Fortune, C. A. (2016). From dynamic risk factors to causal processes: a methodological framework. *Psychology, Crime & Law*, *22*(1-2), 190-202. https://doi.org/10.1080/1068316x.2015.1117080

Ward, T., Polaschek, D., & Beech, A. R. (2006). *Theories of sexual offending*. John Wiley & Sons. https://doi.org/10.1002/9780470713648

Ward, T., & Wilshire, C. E. (2018). Explanation in Forensic Neuroscience. In A. R. Beech, A. J. Carter, R. E. Mann, & P. Rotshtein (Eds.), *The Wiley-Blackwell handbook for* *forensic neuroscience.* Oxford, UK: Wiley-Blackwell. https://doi.org/10.1002/9781118650868

Ward, T., Wilshire, C., & Jackson, L. (2018). The contribution of neuroscience to forensic explanation. *Psychology, Crime & Law*, 1-15. https://doi.org/10.1080/1068316x.2018.1427746

Weisberg, D. S. et al (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*, 470-477. https://doi.org/10.1162/jocn.2008.20.3.470

Yaffe, G. (2016). Mind-reading by brain-reading and criminal responsibility. In Patterson, D., & Pardo, M. S. (Eds.), *Philosophical Foundations of Law and Neuroscience*. (pp. 137-160) Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198743095.003.0008

Yang, Y., & Raine, A. (2009). Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: a meta-analysis. *Psychiatry Research: Neuroimaging*, *174*(2), 81-88. https://doi.org/10.1016/j.pscychresns.2009.03.012