The Neural basis of Social Information Processing: Influencing factors and associated neural mechanisms

Stephanie Julie Wake

PhD

University of York

Psychology

June 2019

Abstract

Understanding social information posits one of the most important leeway's into being a successful member of society. The current doctoral thesis aims to shed light on the complex factors that facilitate and hinder this process, and the underlying neural mechanisms. Presented is an introductory chapter, three empirical chapters with relevant linking chapters, and a conclusive chapter. The first empirical chapter assesses the neural correlates of socio-political information processing across political orientation, in order to understand whether intolerance is accounted for by specific attitudinal orientation or opposing ideology. In this, no neural or behavioural variation is found in the processing of socio-political information across the political left-right, indirectly supporting the notion intolerance stems from ideological conflict. The second empirical chapter assesses the more specific role of the posterior medial frontal cortex (pMFC) in social conflict processing, specifically a role in conflict detection as opposed to resolution via behavioural amendment. Here is found that dorsal medial prefrontal cortex activity (dmPFC) was sensitive to social conflict detection as opposed to conflict resolution. The final empirical chapter assesses whether as humans, we have a specified neural network and circuitry dedicated to social information processing exclusively. Using multivariate analysis techniques, it was found the activation patterns elicited in the ventral striatum were alike between monetary versus social reward. This indicates a subset of neurons responded similarly across both types of reward, signifying a common neural code in some regions for processing both social and non-social information. Collectively this knowledge can aid future research in continuing to decipher the mechanisms relevant to social information processing with more direction. This assists in not only the development of scientific understanding, but also paradigms aiming to produce positive behavioural change within social contexts.

Table of Contents

List of Tables

List of Illustrations

Chapter 1

Chapter 2

Chapter 4

Chapter 6

Acknowledgments

I first of all would like to thank my supervisor, Keise Izuma, for his continuous support and mentorship during the entirety of my PhD. You taught me not only fundamental academic principles that will serve me the rest of my career, but also important life lessons I will always carry. Secondly, I'd like to thank my secondary supervisor Sven Mattys, as well as my thesis advisory panel members Steven Tipper and Beth Jefferies, for both their practical and moral support. All of the academic staff at the University of York, including the neuroimaging centre, present a warm and welcoming environment that help foster personal and technical growth.

Importantly, without the love and support of my parents, Julie and David, none of my success would be possible. I am wholeheartedly grateful for their passion regarding education throughout my life. You always pushed me to be the best I can be and made me believe anything was possible. I'm sure if it wasn't, you both would have found a way to make it possible.

Finally, and unreservedly owed the utmost thanks is my husband, Callum. My love and gratitude for you is eternal. On days when I couldn't believe in myself, it was your unshaken belief coupled with your endless encouragement and love that brought me right back. It's with no doubt that without you I couldn't have done this. Thank you.

Author's declaration

The present thesis contains original work that was completed by the author, Stephanie Wake, under the supervision of Dr Keise Izuma. This work has not previously been presented for an award at this, or any other University. All sources are acknowledged as references.

The original research and data set described in Chapter 4 of the current thesis was supported by the following grant: JSPS London JBUK Japan Award (to K.I.), JSPS KAKENHI Grant Number 15K12777 and 17H00891 (to K.N.) and University Grant for Research Facility and Project, Kochi University of Technology (to R.A. and K.N.).

The data set described in Chapter 6 of the current thesis was supported by Grant-in-Aid for Scientific Research S#17100003 to N.S. from the Japan Society for the Promotion of Science.

The empirical work presented in Chapter 2 in the current thesis has been submitted to the following peer-reviewed journal:.

Wake, S. J., & Izuma, K., (submitted). Neural Correlates of the Left/Right divide: Not Just a Right-Wing-Thing? *NeuroImage.*

The empirical work presented in Chapter 4 in the current thesis has been published in the following peer-reviewed journal:.

Wake, S. J., Aoki, R., Nakahara, K., & Izuma, K., (2019). Elucidating the role of the posterior medial frontal cortex in social conflict processing. *Neuropsychologia,* 132*,* 107124.

The empirical work presented in Chapter 6 in the current thesis has been published in the following peer-reviewed journal:

Wake, S. J., & Izuma, K., (2017). Distinct Neural Representations for Social Versus Monetary Rewards in the Striatum. *SCAN*, 10, 12, 1558-1564.

This work has also been presented at the following conferences:

Wake, S. J., Aoki, R., Nakahara, K., & Izuma, K., (2018). Elucidating the role of the posterior medial frontal cortex in social conflict processing. I presented a poster at the *Frontiers in Social Neuroscience* summer school, 28th June.

Wake, S.J., & Izuma, K. (2016). Distinct Neural Representations for Social Versus Monetary Rewards in the Striatum. I presented a poster at the Oxford Autumn School in Cognitive Neuroscience, 29th-30th September.

For Callum

Chapter 1

Social information processing and the role of social neuroscience

Understanding the social world around us is one of the primary gateways through which we assimilate into society. One must be able to attend to, encode, and understand social cues in order to make sense of, and appropriately respond to, the people and environment around them (Crick & Dodge, 1994; Dodge, 2014). Though this process is seemingly transient across the typically developing population, numerous factors can disrupt this mechanism, and the specific procedures involved in the handling of social information remain still somewhat unresolved. The collaboration of social neuroscience offers additional tools to further investigate the basis of social information processing.

"In fact, neuroscience might offer a reconciliation between biological and psychological approaches to social behaviour in the realisation that its neural regulation reflects both innate, automatic and cognitively impenetrable mechanisms,…" (Adolphs, 2003)

Social neuroscience: a brief overview

Social neuroscience is an interdisciplinary field interested in the association of social and biological factors, specifically the cognition, affect, and behavioural entities with neural, hormonal, cellular, and genetic entities (Cacioppo & Decety, 2012). Neuroscience involves the investigation of the brain and nervous system, whilst social psychology is a broad field comprised of two key elements, the study of intrapersonal level processes i.e. social cognition, and interpersonal/group processes i.e. social interaction (Cacioppo, Berntson, & Decety, 2010). Until the end of the 20th century, biological sciences and social sciences were not seen as mergeable. Social behaviour was vastly studied in the mid to late 1900's, classic work such as Asch's (1952) "line" conformity study, Zimbardo and White's (1972) Stanford

prison experiment (for recent controversy surrounding this, though, see Blum, 2018), and Milgram's (1963) “electric shock” social obedience study (admittedly pushing ethical boundaries) made great progress in understanding some of the fundamental principles of human social behaviour. Nevertheless, these fascinating behaviours were not seen as direct reflections of biological processes, but of somewhat separate and irrelevant to “cellular level function” (Llinás, 1977).

Cacioppo and Decety (2012) succinctly outline three key principles of social neuroscience. The first involves the concept of multiple determinism, which argues that specific occurrences can have multiple attributes and reflections across different levels of processing. For example, the social psychology field has noted in detail social factors predicative of Social Anxiety disorder (for example Carleton, Collimore, & Asmundson, 2010; Clark, Watson, & Mineka, 1994), and the biological/neuroscience field has also outlined several markers relevant for ones predisposition to the disorder (Klumpp, Fitzgerald, & Phan, 2013; Prater, Hosanagar, Klumpp, Angstadt, & Phan, 2013). The contribution of both fields makes for a more complete picture. Second is the notion of nonadditive determinism, arguing that single entities of an overarching mechanism cannot always solely relate to the whole. Meaning, we need to consider all information in conjunction with itself for any clear pattern or relevance to emerge, for example social attitudes and neural correlates. Ignoring the interaction between entities risks crucial mechanisms being missed. Last is the principle of reciprocal determinism, the notion that biological and social entities can reciprocally impact behaviour. For example, hippocampal volume is associated with greater spatial memory (Sherrill, Chrastil, Aselcioglu, Hasselmo, & Stern, 2018), but as a study on London taxi drivers shows (Maguire et al., 2000, see also for evidence of causation 2003), plasticity of the brain means environmental influence (requiring memory of numerous complex routes) significantly contributes to brain structures, in that the hippocampal grey

matter volume of London taxi drivers is significantly larger to that of controls. Without the examination of all available information brought together via social neuroscience, important procedures and validation of social/behavioural models could lack.

Though there exists an abundance of developing neuroimaging techniques widely used in the mapping of social processes to a cognitive and neural basis, the focus of the current thesis will be that of functional magnetic resonance imaging (fMRI) techniques. fMRI was first introduced in the 1990s with early work by Ken Kwong (Kwong et al., 1992) and Seiji Ogawa (Ogawa & Lee, 1990; Ogawa, Lee, Kay, & Tank, 1990; Ogawa, Lee, Nayak, & Glynn, 1990). It operates by identifying the changes in blood oxygenation, coined a blood oxygenation level-dependant (BOLD) response, and flow in reply to neural activity, the basic premise being when a particular brain region is more active, it therefore expends more oxygen thus increases the blood flow to that area. Thereby allowing for functional mapping of brain regions active in response to particular events. fMRI remains one of the most frequent tools used in neuroimaging.

The mechanisms behind social information processing

Relevant to initially outline are the mechanisms currently understood for how we process social information on a rudimentary level. Perhaps first appropriate to consider is Crick and Dodge's (1994) model of Social Information processing, initially based on children, which outlines six steps. Briefly, the first involves encoding of external and internal cues from one’s surroundings. Second, attributions are formed from said cues to determine what motivations underlie the behaviour of others. The third stage involves selecting an objective to govern the preferred outcome, whilst the fourth generates possible responses and actions. The fifth evaluates said responses, assessing suitability for the particular situation, as well as the likelihood of a preferred outcome. Finally, the sixth stage is the tangible behavioural response.

Integrated in a later account of this model is more emphasis on the emotion and affect felt towards who we are interacting with in the first and second stages (see Lemerise & Arsenio, 2000).

A similar and parallel account put forward by Adolphs (2010) conceptualised social information processing into three wide-ranging dimensions: First is perception, this involves the basic input of stimuli from the world around, involving all our senses ranging from visual input such as face perception, auditory perception of speech, and olfactory perception of pheromones. All involve multiple and varied processing streams within the brain. Second is cognition, where we begin to make attributes and inferences regarding any input received in the first stage. This involves more complex processes, importantly, Theory of Mind (ToM) (Premack & Woodruff, 1978). This is the application of mental states (broadly an individual's beliefs, desires, knowledge, emotions, and intentions) to not only oneself but to others, used to understand and predict others' behaviour. Importantly, it is this stage that our cognitive inferences, which are ultimately subjective, may bias information we receive (Adolphs, 2010). To give an example, research on attribution bias shows we tend to infer more favourable evaluations to ourselves and our in-group compared to the out-group (Hewstone, 1990). Third is the regulation of thought, emotion, and actions following the perception and cognition surrounding a particular event, so as to fall loosely in line with one's social norms and values. This could involve external behavioural modifications linked to social conformity (Izuma & Adolphs, 2013), or internal controls on attention to serve goal attainment (Hofmann, Schmeichel, & Baddeley, 2012), for example.

Since social information processing mechanisms deal with particular links amongst surroundings, cognition, and behavioural response, interference within any one stage has costs. For example, misattribution of another's intentions due to bias can lead to further polarization

of attitudes. It is therefore important to gain an in-depth and rich understanding into the elements that can facilitate or hinder these working mechanisms.

Influencing factors: psychological and neural mechanisms

Many factors influence the manner in which social information is perceived, processed, and regulated. These can range from top down influences such as group membership to bottom up influences such as individual traits and social attitudes. Though there is an abundance of factors to consider, focused on here are some key influences relevant to the current thesis.

1.1 Social Attitudes, Stereotypes, and Bias: Psychological accounts

Relevant to consider the impact upon perceiving social information are social attitudes, stereotypes, and bias. Gilbert and Hixon (1991. p. 510) pertinently characterised stereotypes as "*tools that jump out of a* [figurative cognitive] *toolbox when there is a job to be done*". In other words, we utilise a body of information that is oversimplified, overgeneralised, and is quickly and easily accessible to process information about the world with moderate exertion (Allport, 1954; Tajfel, 1969). Bias, commonly linked with stereotypes, is defined as an imbalanced inclination or prejudice towards a specific target. The examination of bias in the social psychology field is by no means a contemporary idea. Research demonstrates its effect vigorously within race relations (e.g. Crisp & Meleady, 2012; Crisp & Turner, 2011; Hutchinson, 2014), gender effects (e.g. Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Wright & Sladden, 2003), age bias (Hills & Lewis, 2011; Levy & Banaji, 2002), plus various other social and minimal intergroup paradigms (e.g. Van Bavel, Packer, & Cunningham, 2008).

A particular form of bias relevant to the current thesis considers the general congruency of social information, coined *confirmation bias*. This typically refers to the pursuit or

construing of evidence that corroborate one's own existing beliefs and world view (see Nickerson, 1998). Indeed, a relative amount of research has demonstrated this effect (Fischer, Greitemeyer, & Frey, 2008; Knobloch-Westerwick, Mothes, Johnson, Westerwick, & Donsbach, 2015; Lord, Ross, & Lepper, 1979), an example of which comes from an experiment conducted in the weeks leading up to the 2012 US presidential election by Knobloch-Westerwick, Johnson, and Westerwick (2015). They found individuals exhibited a robust bias in selective exposure toward attitude-consistent online search results, spending 64% more time on attitude-consistent messages, being especially pronounced among individuals that attached high importance to the issues. Relatedly, Sunstein, Bobadilla-Suarez, Lazzaro, and Sharot (2016) found when assessing the effect climate change stance has on the processing of evidence for and against, climate change deniers were significantly less likely to revise/update their beliefs regarding evidence supporting climate change, and vice versa for pro-climate changers.

Taken with the rising popularity of receiving information from an online forum (Smith, 2013), this particular mechanism is important to assess. The readiness of information available on the internet means exposure to information that reinforces present outlook is arguably easier than ever. Not only does this somewhat stem progressive conversation regarding conflicting views, it also can lead to polarization of views, and ultimately, intolerance (Fernbach, Rogers, Fox, & Sloman, 2013).

1.2 Social Attitudes, Stereotypes, and Bias: Neural accounts

Research has denoted the processing of information that disconfirms one's outlooks is arbitrated by regions involved in error processing and conflict detection, particularly including the anterior cingulate cortex (ACC), adjacent medial frontal regions and lateral prefrontal regions (Botvinick, Cohen, & Carter, 2004; Greening, Finger, & Mitchell, 2011; Sharot, Korn,

& Dolan, 2011). Further, research utilising transcranial magnetic stimulation (TMS; a neuroimaging technique that uses magnetic fields to stimulate or inhibit brain activity) to supress activity in the medial prefrontal cortex found this reduced participants ability to successfully update inconsistent information about others during an impression update task (Ferrari, Vecchi, Todorov, & Cattaneo, 2016). Relatedly, and especially allied to the current thesis, Hughes, Zaki, and Ambady (2017) found participants were more likely to update their impressions regarding negative information during an impression formation task about out-group members, but importantly not in-group members. Less engagement in the dorsal anterior cingulate cortex (dACC), temporoparietal junction, insula, and precuneus when processing negative information about the in-group was found, but again, not for the out-group. This suggests these particular neural structures are potentially important for updating one's impression, specifically when information fits an individual's pre-existing ideas. This allies previous work that outlines the posterior medial frontal cortex (pMFC; encompassing the ACC, see Figure 1.1 for visual aid of key brain labels used throughout the current thesis) as being an important structure during more general social conflict detection (Campbell-Meiklejohn, Bach, Roepstorff, Dolan, & Frith, 2010; Izuma & Adolphs, 2013; Izuma et al., 2010; Wu, Luo, & Feng, 2016).

Offering an additional example utilising a different form of bias, Sharot, Korn, and Dolan (2011) examine the behavioural and neural effect regarding *optimism bias* (the belief more positive as opposed to negative events are likely to happen in the future, thought to be an adaptive mechanism that helps reduce anxiety and depression). Briefly, Sharot and colleagues found that when participants were asked to estimate their likelihood of being involved in certain negative life events (e.g. getting cancer, being involved in a car crash), when faced with the genuine statistics of events happening, participants were significantly less likely to update their beliefs about negative life events happening compared to positive. Interestingly, participants

who scored high for optimistic tendencies showed significantly less activation in the right inferior prefrontal gyrus (IFG) when processing negative information about the future. This is accredited to less efficient encoding by the authors due to the association with tracing negative estimation errors in this region.

Overall, when individuals are particularly motivated (consciously or subconsciously) to maintain a particular world view (be that of optimism or political stance), conflicting information may not be encoded competently, primarily implicating regions within the pMFC such as the ACC that potentially impede any subsequent update.

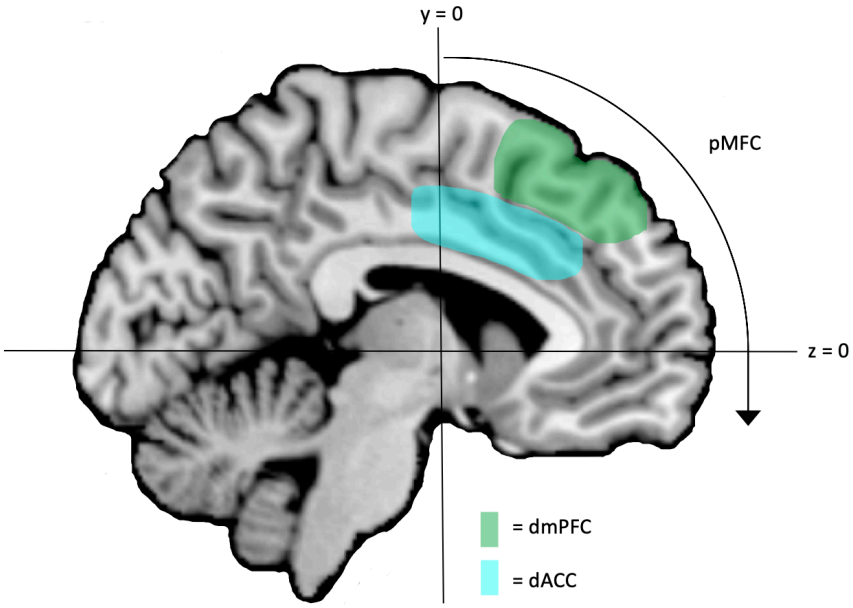


Figure 1.1. Approximate representation of the pMFC, encompassing the dmPFC (green), and the dACC (blue) as referred to within the current thesis.

2.1 Group membership: Psychological accounts

As intensely social species, the notion of group membership is pertinent to consider when deliberating more top down attributes that impact social information processing. Group identities shape not only preferences, viewpoints, and behaviour, but also rudimentary social perception (Van Bavel, Packer, & Cunningham, 2008). For example, and drawing from a similar example as above, people can recognise members of their own race more efficiently (Malpass & Kravitz, 1969; Sporer, 2001). Crisp and Meleady (2012) explain the notion of in-

group favouritism and out-group derogation as being an innate predisposition from early monocultural environments. Those who favoured the in-group and were less likely to engage with out-group members (potentially resulting in intergroup warfare, disease threat, or the loss of mating opportunities) were more likely to survive and reproduce.

Interestingly, aside from more explicit effects of intergroup bias, passive and general perception can also differ between social groups regarding the same information. A classic case study to demonstrate this comes from Hastorf and Cantril (1951), who conducted an experiment following an infamous college football game between Princeton and Dartmouth. Whilst both sides were reported to have repeatedly engaged in vicious and illegal gameplay (resulting in several significant injuries on both sides), the authors were initially intrigued after student newspapers from both universities reported on what appeared to be ostensibly separate games. Following this, data was collected from students attending both universities, to further extract perceptions of the event. Results demonstrated that, even when watching a replay of the game, participants seemed to be "seeing" different games. For example, participants from Princeton saw twice as many infractions from Dartmouth, and Dartmouth saw one third more infractions from Princeton. The authors conclude that there was no such thing as a "single game", but many different games that all exist within individuals.

2.2 Group membership: Neural accounts

Basic neural underpinnings of intergroup perception stem from research showing heightened activity in the fusiform face area while viewing own-race faces during fMRI (Golby, Gabrieli, Chiao, & Eberhardt, 2001), with the strongest effects being from participants with a higher sense of intergroup bias. Golby et al., (2001) determined that own-race biases in fusiform activity were due to superior perceptual expertise with own-race faces. Notably however, this concept has also been replicated in non-racial social groups, and even minimal

groups (particularly interesting is the effect upon assignment to an arbitrary group; Bernstein, Young, & Hugenberg, 2007), demonstrating in-group bias via facial recognition may rely less on perceptual expertise and can be explained by sheer social categorisation.

Regarding the divergence of general perception found by the likes of Hastorf and Cantril (1951), neural correlates have also been found to differ between social groups processing the same information. To display this effect stems from processes regarding intergroup perception, an experiment by Hasson, Malach, and Heeger (2010) demonstrated that when participants viewed the same movie clip from the motion picture *The good, the bad and the ugly*, neural responses were similar both within and across participants. Yet, research shows when individuals view the same stimulus from polarized viewpoints, both perception and associated neural correlates differ.

To illustrate, Molenberghs, Halász, Mattingley, Vanman, and Cunnington (2013) found that after dividing participants into groups, creating a competitive task of who could press a button faster, participants judged their in-group to exhibit quicker response's than their out-group when viewing clips of task performance in an fMRI scanner (even when speeds were matched). Further, this was correlated with an increase in inferior parietal lobule activation (an area associated with action representation; Gallese, Fadiga, Fogassi, & Rizzolatti, 2002) when watching in-group members compared to out-group members. Hence, even when groups are seemingly minimalistic in nature, there still appears to be variation in perception, alongside corresponding neural discrepancies (for a review see Molenberghs, 2013). An additional study by Cikara, Botvinick, and Fiske (2011) scanned avid fans of the Boston Red Sox and New York Yankees whilst viewing a baseball game of the two teams. Adverse outcomes for one's own team activated the ACC (associated with conflict monitoring; Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999) and insula (an area associated with emotion: Cunningham, Raye, & Johnson, 2004, conflict monitoring: Greene, Nystrom, Engell, Darley, & Cohen, 2004; Xiang,

Lohrenz, & Read Montague, 2013: and the detection, regulation, and attentional control of salient stimuli: Menon & Uddin, 2010). Conversely, encouraging outcomes for one's own team activated the ventral striatum (an area heavily associated with reward; Delgado, 2007; Izuma, Saito, & Sadato, 2008; Wake & Izuma, 2017) .

Ultimately it would seem that we experience the actions of our in-group differently to our out-group across a variety of scenarios. Understanding the underlying mechanisms surrounding intergroup perception can give rise to more specific predictors of discrimination, which is important for future designs of interventions aimed to reduce prejudice and discrimination. Importantly, as the research demonstrates more understated groups can provoke diverse perceptions, more subtle intergroup variation such as one's political orientation is important to examine. The current political climate means attitudes are particularly polarized, and so understanding the basic mechanisms amongst more discreet intergroup processes are essential, deciphering underlying cognition to aid in a more well-rounded understanding of social information processing.

A need for depth, clarity, and specificity

All of the above points toward the need for a more precise measurement of the underlying neural correlates of social information processing in order to further our understanding of some fundamental principles. Importantly, the specific mechanisms involved in handling particularly valent social information. This includes social attitudes and political perspective, and how we deal with socially consistent and inconsistent information. Since inconsistent information is related to a lack in subsequent attitudinal/behavioural update (Sharot et al., 2011; Sunstein et al., 2016), it's especially important to investigate these underlying mechanisms. For example, though medial frontal regions are often outlined (Campbell-Meiklejohn et al., 2010; Izuma & Adolphs, 2013; Izuma et al., 2010; Wu et al., 2016), more specific understanding of neural

function is needed to understand if this relates more to conflict detection or impediment of behavioural adjustment. Beginning to understand key aspects associated with the neural response to inconsistent social information in general helps to clarify and build upon cognitive/social models that account for behaviour and information processing.

Additionally important to elucidate social processing mechanisms further is an understanding of whether common versus specified structures are appropriate to understand the principles of social information processing. For example, do we process social information in a special way, with a dedicated system? This being the case, social models drawing from general or non-social research may be less helpful. For instance, the ACC is outlined to be an important structure in general conflict detection as evidence by Stroop/Stroop-like tasks (Barch et al., 2001; Bench et al., 1991; Fan, Flombaum, McCandliss, Thomas, & Posner, 2002; Kerns et al., 2004; Leung, Skudlarski, Gatenby, Peterson, & Gore, 2000), but also in social conflict (Campbell-Meiklejohn et al., 2010; Izuma & Adolphs, 2013; Izuma et al., 2010; Wu et al., 2016). If social information is specialised, the circuitry and involvement of the dACC may be more diverse and/or separate for social stimuli compared to non-social. Understanding if social information is exclusively processed, future research can aim to focus paradigms onto socially pertinent stimuli exclusively to gain a more scrutinised insight into the specific neural structure, function, and circuitry associated.

Therefore, what can be seen from the above literature review is a general need for depth, clarification, and specificity on some unresolved technical questions. Though in some cases the social field has gone into detail to theorise the mechanisms involved, and the neuroscientific field has begun to represent these processes on a neural level, there remains gaps in some of the fundamental aspects.

Thesis Outline and Aim

The current thesis can be thought as overall adding essential knowledge into how humans process particularly valent social information. Within that can be considered two main themes; the first being an in-depth investigation into the neural and cognitive principles associated with predetermined social attitudes and the perception of inconsistent information. Second is the more general idea of a specified social system dedicated to exclusively social information processing. Presented are three empirical chapters in the form of submitted (Chapter 2) and published (Chapter 4 & Chapter 6) manuscripts in peer reviewed journals.

The first empirical chapter (Chapter 2) uses neuroimaging methods to answer a psychological question where the social psychology field may fall short. Fairley recent ongoing debates argue whether the notion of intolerance is predicted via specific attitudinal orientation, particularly politically liberal versus conservative. Past research describes mechanisms typical of conservative orientation such as traditionalism (the desire for previous/past social norms) to be associated with increased levels of intolerance, whereas more recent models describe intolerance as an outcome of more general ideological conflict. To gain an in-depth insight into the variation of intolerance across the socio-political spectrum, the political left versus the political right were assessed in their behavioural and neural responses to opposing political stimuli. This allows us to examine the following concepts:

- Do neural correlates reflective of managing politically inconsistent material shed light on the specific mechanisms involved in socio-political processing.

- Can these neural correlates uncover any tangible disparity between how extremities of the political spectrum (i.e. left versus right wing) process politically inconsistent information, further examining individual differences in socio-political information processing.

The second empirical chapter (Chapter 4) uses neuroimaging methods to assess a neuroscience question, examining the more general neural mechanisms involved in processing

socially conflicting information. The literature outlines the pMFC as a key component in processing conflicting stimuli and the navigation of any behavioural/attitudinal amendment. What is not clear is the specific role of the pMFC within or across these processes. Therefore, a paradigm eliciting cognitive bias was employed so as to dissociate the level of conflict from information more likely to be updated, allowing us to further disentangle the specific role of the pMFC. The group context utilised for this experiment is that of Japan and South Korea, appropriate due to the nature of relations between the countries whom have a long history of political tension (Izuma, Aoki, Shibata, & Nakahara, 2019; Lee, 1985), and now arguably due to high levels of online access, still maintain a distinct disinclination towards each other (see national survey report by *Globe Scan*, 2014). Specifically, Chapter 4 examines:

- What is the more discrete role of the pMFC in processing socially conflicting information.

- Can purposefully designed paradigms separate a role specific to detecting conflict versus the navigation of subsequent behavioural changes, previously conflated in the literature.

The third empirical chapter (Chapter 6) assesses the concept of a specialised neural circuit in humans. What is not fully understood in the literature is if as humans, we have a system and neural network dedicated to the specific and exclusive processing of social information. Since evolution has meant social interaction, therefore social information processing, is imperative in terms of survival and success, considerable research argues for a specialised social system. This concept is studied by examining the neural circuitry for tangible monetary reward in comparison to social reward. This is achieved with a contemporary neuroimaging analysis technique in the form of multivariate-pattern-analysis, which assesses the voxel-by-voxel correlation of activation in specific regions of interest (ROIs) rather than overall univariate strength of activation. This analysis allows for a more detailed overview of the specific neural circuitry involved in any associated event. Specifically, Chapter 6 reanalyses

a previous data set assessing the univariate brain activation for monetary versus social reward, uncovering the following:

- Do humans have a specific social circuitry for processing social reward in comparison to monetary reward.

- Can the use of multivariate analysis techniques further aid our understanding of previously conflated mechanisms.

The questions posed in this research are important to add to the scientific community but are also central socio-cultural questions that as a society, we require information on. Without high level understanding of the mechanisms behind the processing of socially pertinent information, particularly inconsistent or conflicting, interventions are more likely to struggle. Therefore, adding this knowledge to the literature will significantly broaden the scope of not only the impacts of social information processing as a whole, but also adds depth to some technical, unresolved questions.

Chapter 2

Neural Correlates of the Left/Right divide: Not Just a Right-Wing-Thing?

Stephanie J. Wake[1] & Keise Izuma[1,2]

[1] Department of Psychology, University of York, Heslington, York, YO10 5DD, UK.

[2] Department of Psychology, University of Southampton, University Road, Southampton, SO17 1BJ, UK

ABSTRACT

Intolerance is classically argued to be patent among the conservative, political right. However, more recent models suggest intolerance derives from that of ideological conflict rather than specific traits aligned to political orientation. The aim of this study was to observe any variation in the neural correlates using both univariate and multivariate analysis techniques in the processing of politically inconsistent material across the left/right political divide. This was examined by recruiting both left and right wing politically engaged participants and observing their neural responses to politically inconsistent stimuli in an fMRI scanner. Behaviourally we found attitude extremity was positively related to political Intolerance scores across all participants, as well as increased activation in the dorsal medial prefrontal cortex (dmPFC), inferior frontal gyrus (IFG), and thalamus for inconsistent compared to consistent political material. The left insula, dmPFC, superior temporal gyrus (STG), and IFG show more similar patterns of activation for general political material (inconsistent and consistent) compared to apolitical material. Importantly, we found no tangible difference in the processing of politically inconsistent information between our two political groups. This data, though indirectly, supports the Ideological Conflict hypothesis, the notion that intolerance derives from opposing ideology, not specific characteristics of political orientation.

Key Words: Political intolerance, ideological conflict, fMRI, MVPA, insula, dmPFC

INTRODUCTION

In the social and political psychology literature, intolerance (unwillingness to accept views, beliefs, or behaviour that differ from one's own) is typically argued to be most evident among individuals on the political right, as a consequence of psychological needs regarding the maintenance of social norms ("motivated social cognition": Jost, Glaser, Kruglanski, & Sulloway, 2003). The surrounding neuroscience literature has demonstrated variation in both neural structure and function between left versus right wing individuals (Amodio, Jost, Master, & Yee, 2007; Kanai, Feilden, Firth, & Rees, 2011), but there still remains to be seen convincing evidence of both neural *and* cognitive discrepancy in the processing of inconsistent political material across the political divide. More recently, Brandt, Reyna, Chambers, Crawford, and Wetherell, (2014) put forward a model coined the *Ideological Conflict Hypothesis*, which emphasises perceptions that contradicting beliefs are a threat to our own, rather than traditionalism (or other right-wing oriented traits), are key to understanding intolerance (Crawford & Pilanski, 2014; Wetherell, Brandt, & Reyna, 2013). Thus, it is important to better understand the neural correlates of managing inconsistent political material across both political groups in order to uncover the true predictors of intolerance.

Social psychology has outlined classic traits representative of the political left and right. Liberals are argued to score higher on levels of novelty seeking (Jost, Federico, & Napier, 2009), openness to experience and cognitive flexibility/ability (Adorno et al., 1950; Crisp & Meleady, 2012; Jost et al., 2003; Kemmelmeier, 2008), and conservatives tend to demonstrate an increased sensitivity to threat, creating stronger desires for traditional norms/familiarity (Adorno et al., 1950; Jost et al., 2003). Yet, although the left may be more likely to accept norm challenging views and the right may be more likely to reject them- this doesn't necessarily mean there exists the same skew of intolerance to directly opposing/inconsistent political material. For example, Crawford and Pilanski (2014) utilised a least-liked group paradigm (a

method whereby participants choose their least liked social group in order to assess levels of intolerance, rather than the same set of groups i.e. homosexuals, immigrants) and found no difference in political intolerance between US liberals and conservatives. This would suggest the cognitive process behind the managing of politically inconsistent material may actually be similar across the political groups. Since there exists at least several levels of organisation in the processing of complex social information, and only the end goal reaches any tangible measurement via behavioural analysis (Wilson & Bar-Anan, 2008), more in depth assessment may be required to elucidate further this question. The use of neuroimaging to conceptualise these types of effects help shed light on the underlying neural correlates and therefore more discrete components of social information processing. Within the current context this allows for a deeper insight into the specific processes involved in political intolerance, and any discrepancies between the political left and right.

To begin to review brain regions associated with the processing of political material in general, an fMRI study by Zamboni et al., (2009) provides an interesting insight by having participants view both liberal and conservative statements in the scanner. They found that the processing of conservative statements was associated with greater activity in the dorsolateral prefrontal cortex (dlPFC). The authors speculate this could be due to i) the involvement of this area in complex moral decision making between self-interest and fairness (i.e. Cunningham & Zelazo, 2007; Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006), or ii) a result of liberal responses to conservative statements- as the majority of participants were of a liberal/moderate ideology (73.1%). In order to gain a more balanced view of brain activity for directly opposing material, Kaplan, Freedman, and Iacoboni (2007) recruited 10 individuals who were registered Democrats and 10 individuals who were registered Republican during the 2004 United States presidential election, presenting them with faces from one's own political party and faces from an opposing political party. They found increased activity in the dlPFC, anterior cingulate

cortex (ACC) and insula in response to viewing opposing political faces compared to faces from one's own political party. These regions were also positivity correlated with the reported emotional feelings of the participants towards the associated candidates. The neural response is interpreted as activating areas associated with cognitive control (dlPFC, ACC) and emotion (insula), evidencing participants attempt at regulation when viewing opposing political material.

Furthermore, Knutson, Wood, Spampinato, and Grafman, (2006) found when showing participants images of political party leaders (democrats and republicans) faces along with positive/negative words, ventromedial anterior prefrontal cortex (inferior frontal gyrus, ACC, left precentral, superior parietal lobe, and dlPFC) were more active for incongruent (dependant on participants political adherence) compared to congruent and control trials. Brain activity in the frontopolar region was positively correlated with implicit bias and strength of feeling toward the politician, but strength of affiliation toward political party was negatively correlated with the lateral PFC (lPFC). The authors interpret this as a duel response for dealing with political information, one for processing more emotional and stereotypic information regarding opposing political material (i.e. vmPFC), and one for more reflective and deliberate processing (i.e. anterior prefrontal cortex). However, one issue with mPFC activation may be that it represents more a prediction error/conflict, which the mPFC is known to be sensitive to (Campbell-Meiklejohn et al., 2010; Izuma & Adolphs, 2013; Sambrook & Goslin, 2015), rather than more specifically processing inconsistent political stimuli.

Additionally examining participants committed to either republican or democratic candidates in the lead up to the 2004 US presidential election, Westen, Blagov, Harenski, Kilts, and Hamann (2006) demonstrated the motivated reasoning involved in defending one's own political affiliation was associated with the vmPFC, ACC, posterior cingulate cortex (PCC), insula, and lateral orbital cortex. Moreover, it has also been shown that when liberal

participants were presented with counter evidence aimed to change their political beliefs, more belief-countering trials were associated with dmPFC activity, and the most likely to change their mind presented significantly less blood-oxygen level dependant (BOLD) signal in the insula. Thereby suggesting those with more concrete attitudes were more sensitive to counter evidence (inconsistent political material) within the insula (Kaplan, Gimbel, & Harris, 2016).

Although the research discussed provides great impact in bridging the gap between social psychology and *political neuroscience* (a term coined by Jost, Nam, Amodio, & Van Bavel, 2014), the area still lacks an investigation into not only the neural basis of socio-political attitudes, but handling of strictly opposing ideology. One major limitation in the previous studies is that they heavily relied on reverse inference when interpreting brain activations. This can prove particularly problematic when examining functionally heterogeneous structures such as the insula and dACC, both of which were often reported in the previous studies (for a review see Poldrack, 2011).

In order to address the limitation in the present study, we directly compare neural responses to politically inconsistent statements with 1) merely negative and 2) immoral statements using multivariate pattern analysis (MVPA). One idea is that individuals may perceive politically opposing ideas as simply negative, resulting in similar neural responses. The second idea is that politically opposing ideas may be perceived as immoral or *morally disgusting* (Chapman & Anderson, 2013). Research has highlighted moral transgression to be highly predictive of general intolerance (e.g. see Wright, McWhite, & Grandjean, 2014), demonstrated to be more predictive of intolerance than non-moral forms of diversity (Wright, Cullum, & Schwab, 2008). Furthermore, to gain a better insight into psychological reactions to politically opposing ideas, when comparing neural responses to politically inconsistent statements with negative and immoral statements, we not only compare the strength of activation in a region of interest (ROI), but also compare activity pattern across multiple voxels

within a ROI using MVPA. A number of past MVPA studies have demonstrated that different stimuli activated the same brain regions, but activation patterns were different across two conditions, indicating that the two stimuli were supported by distinct neural (or psychological) mechanisms (Haxby et al., 2001; Wake & Izuma, 2017; Woo et al., 2014).

A region heavily associated with disgust or negative emotion in general is the anterior insula (for a review, see Chapman & Anderson, 2013), and research has indicated some overlap in this region in terms of moral transgression/conflict (Greene et al., 2004). As stated above, the insula has often been reported in past political neuroscience studies especially when participants were confronted with contrasting political material, such as opposing candidate faces (Kaplan et al., 2007) and belief-countering stimuli (Kaplan et al., 2016). Though the insula is too a large and functionally heterogeneous structure (i.e. associations are made with: anger and fear, Damasio et al., 2000; anxiety, Critchley, Wiens, Rotshtein, Öhman, & Dolan, 2004; and pain, Peyron, Laurent, & Garcia-Larrea, 2000), it seems the association with socially emotive and both morally *and* politically inconsistent stimuli make this a rational approach for assisting in localising a ROI for comparing ideologically opposing/inconsistent stimuli.

Therefore, the present study aimed to examine levels of political intolerance directly across left wing and right wing participants, comparing neural responses to politically inconsistent statements with negative and immoral statements in specific, predetermined ROIs, i.e. the bilateral anterior insula. This, alongside the use of opposing/inconsistent political stimuli equally across both political groups, generates a more balanced insight into the genuine predictors of intolerance that so far only the social field has begun to make use of.

Accordingly, our hypothesis predicts i) stronger univariate activation within respective ROIs (i.e. insula) for both politically inconsistent statements and immoral and/or negative statements relative to politically consistent statements, ii) similar univariate activation as well as neural pattern identified via MVPA within ROIs for politically inconsistent trials across both

left wing and right wing participants, alongside no significant difference in average political intolerance scores.

METHOD

Participants

Forty healthy university students who possessed strong political attitudes (either liberal or conservative) were recruited for the study. 7 out of 40 were recruited from political societies on the University of York campus (*York Tories, Labour Society, Socialist Society,* and the *UKIP Society*). Additionally, an online questionnaire was distributed amongst students, and participants that scored above a criteria cut off were also invited to participate (participant's needed to indicate on a five-point scale they were "1= Very Liberal", "2= Liberal", or "4 = Conservative", "5= Very Conservative", alongside a score of at least "5" on a ten-point scale assessing strength of political attitudes, a score of "3" on a four-point scale assessing amount of times politics is discussed, and have indicated they have taken part in at least one political activity i.e. signing a petition). The following five participants were excluded from the analysis; One participant was excluded due to excessive head motion (>3mm), three were excluded for providing a moderate orientation rating on the day of the experiment (i.e. did not indicate being liberal or conservative), and a further participant was excluded due to inconsistent answers provided in the scanning session (indicating the participant was not/had stopped paying attention to the stimuli). The final sample consisted of 35 participants, 23 left wing orientated (12 female, mean age = 20.8) and 12 right wing orientated (5 female, mean age = 20.3). All participants gave written informed consent for participation, and the study was approved by the Research and Ethics Committee of York Neuroimaging Centre. It's also important to note the data was collected between November 2016 to April 2018, during and immediately after

the election of President Donald Trump and Britain's leaving the European union, a highly politically relevant period.

Procedure & Task

In the fMRI session, participants were asked to view a series of both political and apolitical statements/scenarios, alongside details of a picture of the individual responsible/involved (see Figure 2.1). There were five experimental conditions; 1) pro-right wing oriented condition, 2) pro-left wing oriented condition, 3) immoral condition, 4) negative condition, and 5) neutral (control) condition, with 24 statements included in each condition (a total of 120 statements). The politically charged stimuli were comprised of 24 pro-right wing orientated statements (e.g., Samantha, age 25, believes the rich are too highly taxed), and 24 pro-left wing orientated statements (e.g., Alison, age 41, believes everybody should receive free health care). Additionally, 24 immoral or "moral disgust" statements were included (e.g. Oscar, age 22, and a group of his friends trip an old man and laugh), as well as 24 negative, non-moral statements (e.g., Joe, age 30, is forced to let the vet euthanize his terminally ill horse). Finally, 24 emotionally neutral control statements (e.g. Sam, age 30, sharpens his pencils ready to sketch a picture of a landscape) were also included.

We included the immoral and negative conditions to compare the responses to politically inconsistent material (political intolerance) with immoral material (moral disgust) and negative material. Alongside the statements, participants were provided with the name, age, and an image of the individual (see Figure 2.1). This was to make material more authentic and ecologically valid to participants. Before the fMRI task, participants were led to believe that each statement was actually mentioned by each individual or reflected something that actually happened to them (in reality, the individuals/statements were created for the purpose of the experiment).

All political statements were selected based on a pilot study (n = 63, 38 female, mean age = 20). Participants were asked to complete UK adapted versions of the political statements used in Zamboni et al., (2009) and rate additional political statements (total n = 128) using a 7-point scale (1 = strongly disagree, 7 = strongly agree). We selected the 88 political statements that were most highly correlated with participants' political orientation. Participants also rated the images of faces for trustworthiness and attractiveness using a 7-point scale (1= extremely untrustworthy/unattractive, 7= extremely trustworthy/attractive) and rated the Negative and Immoral statements for valence (1= strongly negative, 7= strongly positive), which was used to match across conditions. The combination between faces and statements was counterbalanced across participants for the main fMRI experiment.

During each fMRI run, participants were asked to pay attention to each statement and further told they would be asked questions at random time points (to ensure attention is maintained throughout the experiment). Questions were presented, on average, once per 3 trials, and there were 3 types of questions; 1) "How positive/negative did you feel about the previous statement?", 2) "How empathetic did you feel toward the previous person?", and 3) "How strongly did the previous statement make you feel?". Participants had a button box in which they could rate on a scale of 1-3 (1 indicating a lower response and 3 a higher response on all question trials).

Participants took part in four fMRI runs, each lasting 7.2 minutes, viewing a total of 30 statements in each run. Each statement was presented for 6 seconds followed by an inter-trial interval (ITI; 4, 6 or 8 seconds, average = 6 seconds). In each run, 9 questions were randomly inserted. For question trials, the question was presented immediately after the statement presentation, and it remained on the screen for 6 seconds followed by the ITI (see Figure 2.1). Trial order was fixed for all participants (note however that we analysed the fMRI data based on whether each political statement is consistent or inconsistent with participant's political

orientation [see below] so that the order of politically consistent/inconsistent trials was different across the two groups of participants). Additionally, some of the participants also took part in a fifth fMRI run (a 13.3 minute video of various political figures), but this task was not related to the current study and the results are not conveyed within this report.

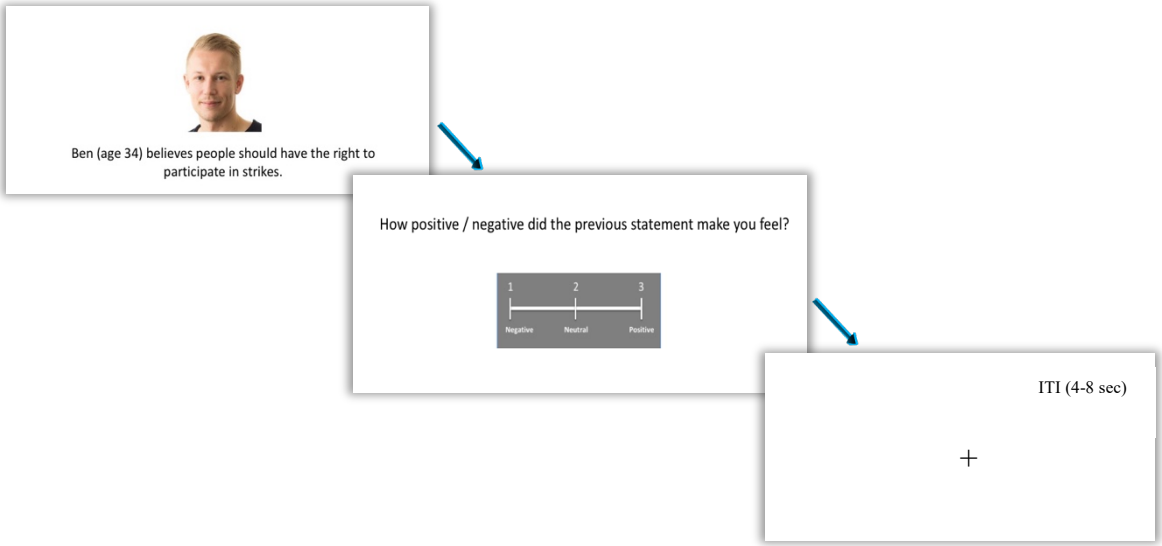


Figure 2.1. Example of a pro-left wing trial followed by a question utilised for fMRI stimuli, as seen by participants inside the scanner.

After the fMRI session, participants completed 2 questionnaires. First, a Political Knowledge and Interest questionnaire, including 7 "true or false" Political Knowledge items (adapted from Larcinese, 2007; i.e. Margret Thatcher was a conservative prime minister: *true or false*), and 3 items using a 4-point Likert scale measuring self-reported knowledge, interest, and amount of time politics is discussed (adapted from the Audit of Political Engagement, 2005). Secondly, participants completed a Political Intolerance questionnaire, comprised of 11 items (adapted from Crawford & Pilanski, 2013). For this measure, a least-liked group paradigm was utilised. This involves participants giving ratings on a group specifically opposing to them, rather than rating the same groups across participants (i.e. left wing person rating the same groups as a right wing person). The questionnaire used by Crawford and Pilanski (2012) was adapted to UK equivalents, i.e. "I think that the Democratic (*Republican*)

Party should not be allowed to visit college campuses in order to register potential voters" was changed to "I think that the Communist Party of Britain (*Britain First Party*) should not be allowed to visit university campuses in order to register potential voters" (right wing equivalents that were distributed to left wing participants are provided in parenthesis). This provides a measure of intolerance towards opposing ideological view- rather than the same set of social groups, reducing the likelihood of measurement bias (for full set of measures used, see Appendix 1a & b). Finally, general demographics including Social and Economic orientation scores (7-point Likert scale; 1= strongly liberal, 7= strongly conservative) were collected. Upon completing the experiment, all participants were debriefed, thanked, and paid £25 or given equivalent course credits.

fMRI Data Acquisition

Images were acquired using a GE Signa 3T MRI system at York Neuroimaging Centre. For functional imaging during the sessions, interleaved T2*-weighted gradient-echo echo-planar imaging (EPI) sequences were used to produce 38 continuous 3mm thick trans axial slices covering the entire cerebrum and cerebellum (repetition time [TR] = 3000ms; echo time [TE] = 30ms; flip angle [FA] = 90; field of view [FOV] = 288mm; voxel dimensions = 3.0 × 3.0 × 3.0 mm). A high-resolution anatomical T1-weighted image (38 continuous 3mm thick trans axial slices covering the entire cerebrum and cerebellum; 512x512 matrix over a 288mm FOV, voxel dimensions = 0.56x0.56x3mm) was also acquired for each participant.

fMRI Data Pre-processing

The fMRI data was analysed using SPM12 (Welcome Department of Imaging Neuroscience) implemented in Matlab (MathWorks). The first four volumes were discarded to allow for T1 equilibration. Head motion was corrected using the realignment program in

SPM12. Following realignment, the volumes were normalised to MNI space using a transformation matrix obtained from the normalisation of the first T1 image of each individual subject to the template T1 image, and then applied to all EPI images. The normalised fMRI data were spatially smoothed with a Gaussian kernel of 8 mm (full-width at half-maximum) in the x, y, and z axes.

fMRI Data Analysis

A first level analysis using a general linear model (GLM) was run with the intention to identify brain regions activated in response to politically inconsistent statements. Data was analysed based on the five following conditions; 1) Consistent (politically charged statements consistent with participants political orientation), 2) Inconsistent (politically charged statements inconsistent with participants political orientation), 3) Immoral, 4) Negative, and 5) Control. Entered into the model was: 1) presentation of Consistent trials (duration = 6 seconds), 2) presentation of Inconsistent trials (duration = 6 seconds), 3) presentation of Immoral trials (duration = 6 seconds), 4) presentation of Negative trials (duration = 6 seconds), and 5) presentation of Control trials (duration = 6 seconds). Other regressors that were of no interest, such as the Question trials, six motion parameters, the session effect, and high-pass filtering (128 sec) were also included.

The first level GLM analysis yielded the following ten main contrast images for each participant, which were submitted to group level analyses (i.e., one-sample t-test) and MVPA (see below); 1) Consistent, 2) Inconsistent, 3) Immoral, 4) Negative, 5) Control, 6) Consistent > Control, 7) Inconsistent > Control, 8) Immoral > Control, 9) Consistent + Inconsistent + Negative + Immoral > 4Control (localiser contrast), and 10) Inconsistent > Consistent. In addition to the group analysis including all 35 participants, we ran another group analysis directly comparing the two groups of participants (left wing participants [n = 23] vs. right wing

participants [n = 12]).

Correlation-based MVPA

In order to gain a closer insight into the sensitivity of brain activity in response to politically Inconsistent stimuli, we also ran a multivariate pattern analysis (MVPA). This calculated a voxel-by-voxel Pearson's correlation coefficients of activation for all five conditions (versus implicit rest) within each ROI for each participant. Correlational values were then fisher-z transformed and submitted to group level analysis (i.e. one sample t-test, or Independent sample t-test to compare between left and right wing participants). To reduce any risk of potential outlier bias, we also ran the same MVPA analyses using Spearman's correlation, but it produced virtually the same results.

Generating ROIs

To isolate ROIs for both the univariate and MVPA analyses, we applied two approaches. Firstly, a statistical map signifying brain regions associated with negative affect yielded from a meta-analysis by Lindquist, Satpute, Wager, Weber, and Barrett, (2015) was applied to our localiser contrast image Consistent + Inconsistent + Negative + Immoral > 4Control. We included all experimental trials excluding Control so as to avoid manufacturing bias when comparing our ROIs between conditions (i.e. "double dipping"). Clusters which survived the set threshold (height $p < 0.001$ uncorrected, and cluster $p < 0.05$ corrected for family-wise-error: FWE) were then defined as a respective ROI, from which four were created; the dorsal medial prefrontal cortex (dmPFC: x = -10 y = 48 z = 30), the left superior temporal gyrus (STG: x = -50 y = -60 z = 22), the left inferior frontal gyrus (IFG: x = -54 y = 18 z = 12), and the thalamus (x = -8 y = -6 z = 4) (see Figure 2 for axial slices of all ROIs included in analysis). Contrary to our expectation, we didn't find any insula activation with the localiser contrast. To

explore how the insula responded to each type of statement, we defined the insula ROI utilising Neurosynth (http://www.neurosynth.org/). We extracted the peak coordinates of both left and right insula activation provided by a term-based meta-analysis of 84 studies applying the term "negative emotional" (left insula: x = -34 y = 18 z = -2; right insula: x = 38 y = 20 z = -4).

The beta values for each insula ROI were extracted via an 8mm sphere centred around each of the given coordinates, and the beta values for our functionally defined ROIs were extracted from the full clusters (left insula: 257 voxels, right insula: 257 voxels, dmPFC: 1316 voxels, STG: 645 voxels, IFG: 963 voxels, thalamus: 340 voxels; see Figure 2.2). To compare both across group and between groups response to the experimental stimuli, average beta values were extracted from all ROIs from the group contrast images Inconsistent > Control, Consistent > Control, Immoral > Control, and Negative > Control. This was also applied for left and right wing participants separately, additionally including the Inconsistent > Consistent contrast. To control for multiple comparisons in our univariate analysis, all post hoc tests and t-tests utilised Bonferroni-Holm corrected *p* values.

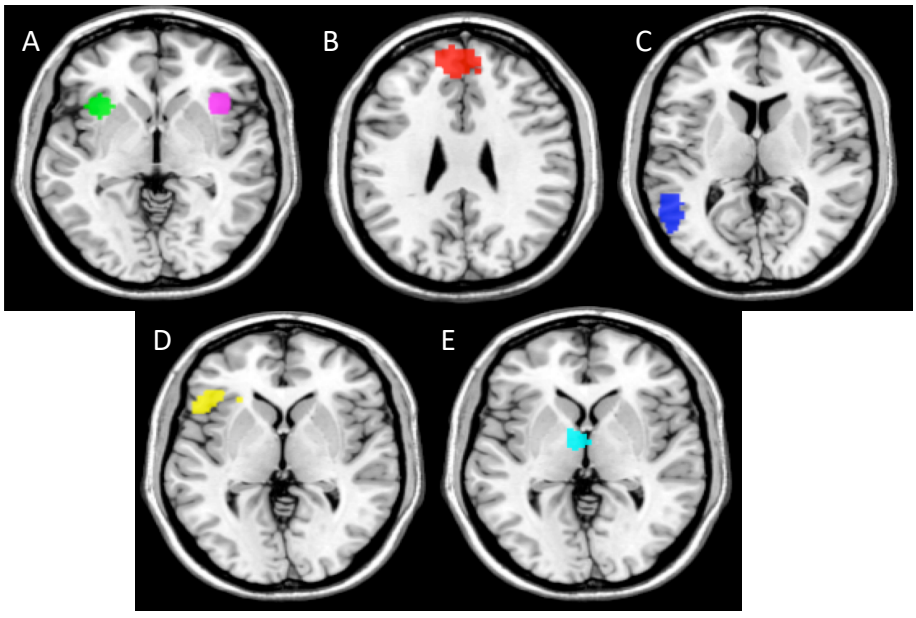


Figure 2.2. (**A**) Axial slice (z = -2) showing ROIs identified via Neurosynth, green signifies the left insula, violet signifies the right insula. (**B**). Axial slice (z = 27) showing functionally defined ROI dmPFC. (**C**) Axial slice (z = 10) showing functionally defined ROI STG. (**D**) Axial slice (z = 1) showing functionally defined ROI IFG (**E**). Axial slice (z = 1) showing functionally defined ROI thalamus.

Behavioural Data Analysis

Participants Political Knowledge measure was calculated by totalling the number of correct responses given. Participants Intolerance ratings were reversed scored where appropriate and averaged, producing a mean Intolerance score for each participant (Cronbach's Alpha = 0.79). Social and Economic Orientation scores were also averaged in order to produce general Orientation scores for each participant. Furthermore, in order to assess strength of attitude, all left wing Orientation scores (Orientation, and Social and Economic Orientation separately) were reversed scored so as to directly compare with right wing scores, a higher score representing a more extreme/strong orientation.

RESULTS

Behavioural Results

Political Intolerance and Orientation

Interestingly, a two sample t-test demonstrated significantly higher Intolerance scores for left wing participants (a higher score indicates greater intolerance) versus right wing participants ($t(33) = 2.15$, $p = 0.04$, Cohen's $d = 0.77$) (see Figure 2.3A). Additionally, there was a significant negative correlation between participants Orientation scores (no reverse scoring applied so a lower score represents more liberal views, and a higher score represents more conservative views) and political Intolerance scores ($r = -0.37$, $p = 0.03$), suggesting the more conservative participants are, the less politically intolerant (see Figure 2.3B).

Following up the Orientation score analysis, a two sample t-test demonstrated significantly higher Orientation scores (reversed scored so a higher score represents stronger orientation) from the left wing participants compared to the right ($t(33) = 6.40$, $p < 0.001$).

To further examine if specifically left wing attitudes were associated with increased Intolerance, or simply stronger attitudes were, we ran a Pearson's correlation analysis and

found a significant positive correlation between the (left wing reversed) Orientation scores and Intolerance scores ($r = 0.44$, $p = 0.009$). This suggests that the increased Intolerance observed for left wing participants compared to the right is more likely due to generally stronger attitudes/affiliations than specific left wing characteristics (see Figure 2.3C).

Finally, we examined the relationship between social versus economic values across each political group. We conducted a 2 (Political Group: left wing vs. right wing) × 2 (Orientation Type: Social vs. Economic) mixed ANOVA, and found a significant main between-subjects effect of Political Group ($F(1,33)= 40.87$, $p < 0.001$, $\eta p^2 = 0.55$), a significant interaction effect ($F(1,33)= 18.13$, $p < 0.001$, $\eta p^2 = 0.36$), but no main within-subject effect of Orientation Type ($p= 0.12$). Intriguingly, our results demonstrate the direction of effect is opposing between the groups (see Figure 2.3D). The left wing group produced stronger ratings for Social Orientations, whereas the right wing group produced stronger ratings for Economic Orientations.

Overall, this indicates the left wing participant's are more extreme in their overall political opinions and Intolerance scores than are our right wing sample. Moreover, we speculate that the left wing sample place more emphasis on social values, whereas our right wing sample place more emphasis on economic values.

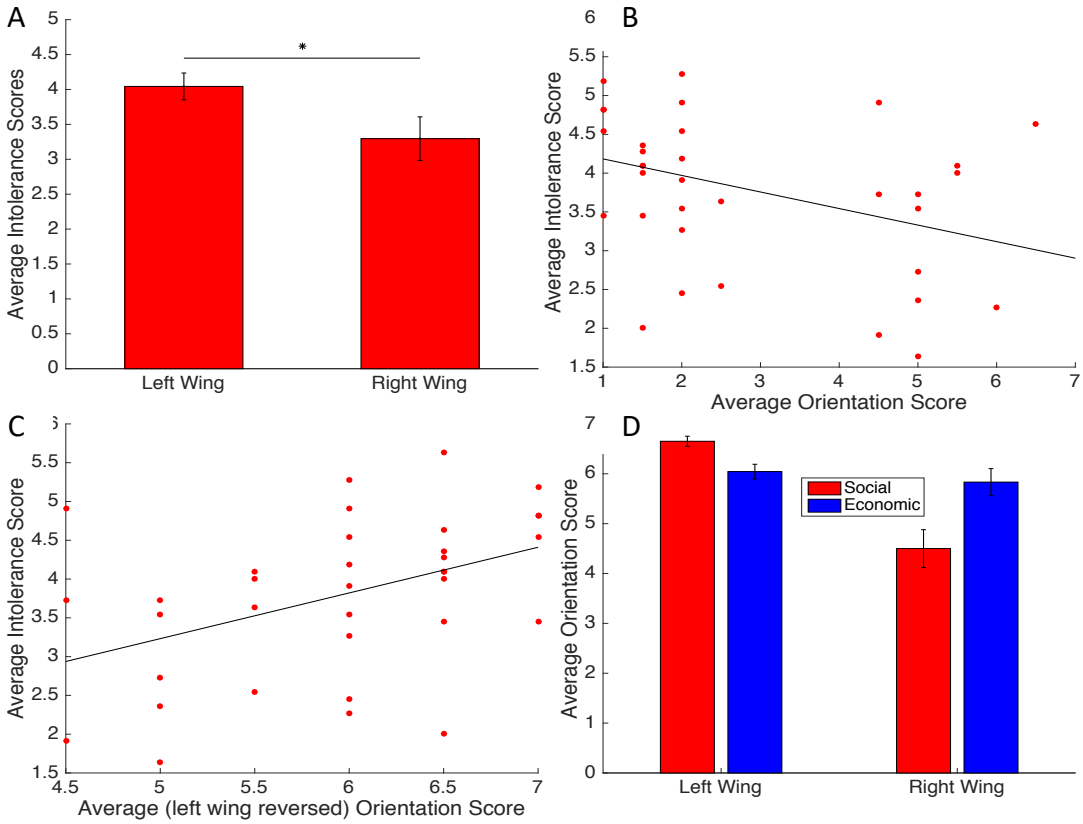


Figure 2.3. (**A**). Bars represent mean Intolerance scores for left wing and right wing participants. Error bars denote standard error of mean (SEM) (**B**). Scatter plot demonstrating negative correlation between participants' Orientation and Intolerance score. (**C**). Scatter plot demonstrating positive correlation between participants' (left wing reversed) Orientation and Intolerance score. (**D**). Bars represent mean social and economic orientation scores for left wing and right wing participants. Error bars denote SEM.

Political Knowledge

The average number of correct answers across participants was 4.69 (approximately 67%). We found no significant differences in the number of correct answers between the left wing participants versus the right wing participants ($p = 0.13$). The number of correct answers was also not significantly related to Intolerance scores ($p = 0.21$), or (the left wing reversed) Orientation scores ($p = 0.19$).

Self-reported Political Knowledge, Political Interest, and Political Discussion

An independent samples t-test found no significant differences between left and right wing participants for Self-reported political Knowledge and Political Interest, but left wing participants conveyed significantly higher rates of self-reported frequency of Political Discussion compared to right wing participants (see Table 2.1 for associated descriptive and inferential statistics).

Table 2.1. Descriptive and Inferential Statistics of control measures; self-reported political knowledge, political interest, and political discussion.

Measure	Total mean (*N*=35)	Left Wing mean (*N*=23)	Right Wing mean (*N*=12)	*t* statistic	*p* value
Political Knowledge (self-report)	3.11 SD = 0.58	3.17 SD = 0.58	3.00 SD = 0.60	0.83	0.41
Political Interest	3.51 SD = 0.56	3.61 SD = 0.50	3.33 SD = 0.65	1.39	0.17
Frequency of Political discussion	3.71 SD = 0.62	3.91 SD = 0.29	3.33 SD = 0.89	2.89	0.006**

SD = standard deviation, *= $p < 0.05$, **= $p < 0.01$.

Furthermore, across all participants (n = 35), Intolerance scores were significantly correlated with frequency of political discussion ($r = 0.34$, $p = 0.04$), but were not correlated with self-reported political knowledge ($p = 0.80$), or political interest ($p = 0.20$). These results suggest left wing participants openly discuss politics more than right wing participants, and this frequency of discussion is related to higher levels of political Intolerance across participants.

fMRI Results: Univariate analysis

Left Insula

A 2 (political group: Left vs Right) × 4 (condition: Consistent, Inconsistent, Immoral, and Negative) mixed ANOVA revealed a significant main effect within-subjects for condition ($F(3,99) = 3.98$, $p = 0.01$), but no between-subject main effect of political group ($p = 0.46$), or interaction effect ($p = 0.46$). Post hoc tests revealed only a significant difference between Inconsistent versus Negative ($t(34) = 3.96$, $p = 0.002$), with no other differences between conditions (Figure 2.4A).

Considering this result, we then combined the data between groups for further analysis. One-sample t-tests revealed the average beta values for the Inconsistent condition were significantly different from zero ($t(34) = 4.66$, $p < 0.001$), but the other three conditions were not significantly different from zero (all $ps > 0.31$) (see Figure 2.4A).

Right Insula

We conducted the same 2 (political group: Left vs Right) × 4 (condition: Consistent, Inconsistent, Immoral, and Negative) mixed ANOVA for the right insula, and it revealed a significant main effect of condition ($F(3,99) = 3.60$, $p = 0.02$), but no main effect of political group ($p= 0.44$), or interaction effect ($p = 0.92$). Post hoc tests again revealed only a significant difference between the Inconsistent versus Negative conditions ($t(34) = 3.25$, $p = 0.02$; Figure 2.4B).

We combined the data between the two groups, and one-sample t-tests revealed that none of the average beta values were significantly different from zero (Inconsistent $p = 0.06$; Consistent $p = 1$; Immoral $p = 1$; Negative $p = 1$) (see Figure 2.4B).

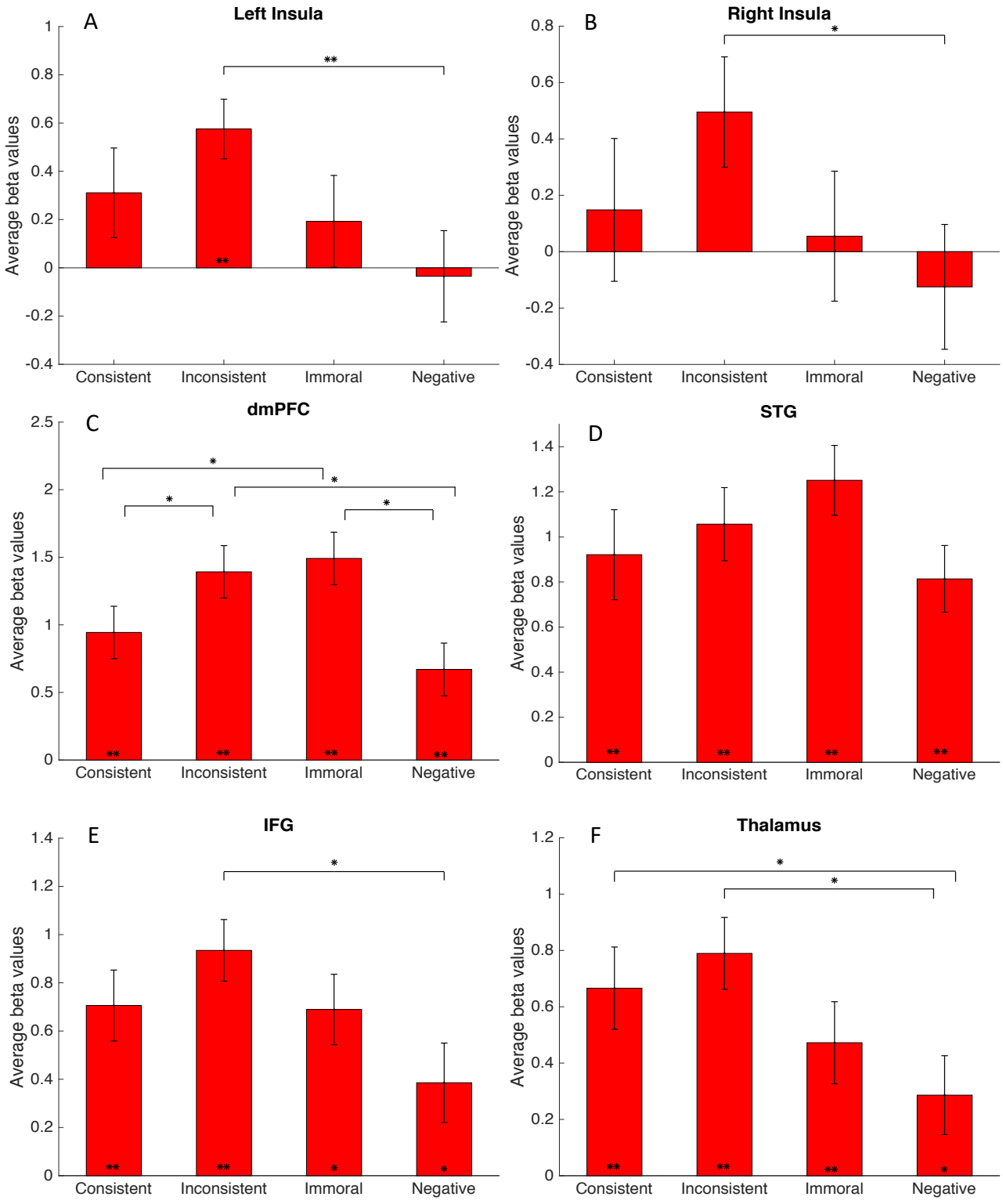


Figure 2.4. Bars represent average beta values for all experimental conditions > control in all relevant ROIs, * $p < 0.05$, ** $p < 0.01$, below asterisks refers to one-sample t-test, above refers to paired t-tests. Error bars denote SEM.

dmPFC

A 2 (political group: Left vs Right) × 4 (condition: Consistent, Inconsistent, Immoral, and Negative) mixed ANOVA for the dmPFC revealed a significant main effect of condition ($F(3,99) = 6.85$, $p < 0.001$), but no main effect of political group ($p = 0.30$), or interaction effect ($p = 0.36$). Post hoc tests reveal a significant difference between Consistent versus Inconsistent ($t(34) = -2.76$, $p = 0.04$), Consistent versus Immoral ($t(34) = -2.76$, $p = 0.04$), Inconsistent versus Negative ($t(34) = 3.38$, $p = 0.01$), and Immoral versus Negative ($t(34) = 4.51$, $p < 0.001$). There were no significant differences between Consistent versus Negative ($p = 0.38$), or Inconsistent versus Immoral ($p = 0.62$; see Figure 2.4C). This demonstrates that Inconsistent and Immoral stimuli activated the dmPFC at a similar degree, both more strongly than Consistent and Negative stimuli, falling in line with the studies hypothesis (stronger univariate activation for Inconsistent and Immoral material within respective ROIs, both relative to Consistent material).

We combined the data between groups, and one-sample t-tests revealed that the average beta values for all four conditions (Consistent, Inconsistent, Immoral, and Negative) were significantly different from zero (all $ps < 0.001$) (see Figure 2.4C).

STG

A 2 (political group: Left vs Right) × 4 (condition: Consistent, Inconsistent, Immoral, and Negative) mixed ANOVA was conducted for the STG, and it revealed no significant main effect for condition ($p= 0.09$), no main effect of political group ($p = 0.45$), or interaction effect ($p = 0.07$).

We combined the data between groups, and a one-sample t-test revealed that the average beta values for all four conditions (Consistent, Inconsistent, Immoral, and Negative) were significantly different from zero (all $ps < 0.001$) (see Figure 2.4D).

IFG

A 2 (political group: Left vs Right) × 4 (condition: Consistent, Inconsistent, Immoral, and Negative) mixed ANOVA for the IFG revealed a significant main effect of condition ($F(3,99) = 3.46$, $p = 0.02$), but no main effect of political group ($p = 0.37$), or interaction effect ($p = 0.13$). Post hoc tests revealed a significant difference between the Inconsistent versus Negative condition ($t(34) = 3.14$, $p = 0.02$), but no other differences between conditions (all ps > 0.05).

We combined the data between groups, and a one-sample t-test revealed that the average beta values for all four conditions (Consistent, Inconsistent, Immoral, and Negative) were significantly different from zero (Consistent $p < 0.001$, Inconsistent $p < 0.001$, Immoral $p = 0.01$, Negative $p = 0.048$, see Figure 2.4E).

Thalamus

A 2 (political group: Left vs Right) × 4 (condition: Consistent, Inconsistent, Immoral, and Negative) mixed ANOVA for the thalamus revealed a significant main effect of condition ($F(3,99) = 5.87$, $p < 0.001$), but no main effect of political group ($p = 0.99$), or interaction effect ($p = 0.31$). Post hoc tests reveal a significant difference between the Consistent versus Negative condition ($t(34) = 3.15$, $p = 0.02$), and the Inconsistent versus Negative condition ($t(34) = 3.39$, $p = 0.01$), but no other differences between conditions (all ps > 0.05; see Figure 2.4F).

We combined the data between groups, and one-sample t-tests revealed that the average beta values for all four conditions (Consistent, Inconsistent, Immoral, and Negative) were significantly different from zero (Consistent, Inconsistent, and Immoral $ps < 0.001$, for the Negative condition $p = 0.03$, see Figure 2.4F).

Whole brain analysis

Finally, in order to more broadly identify any further regions specifically associated with viewing Inconsistent political statements compared to Consistent statements, we examined the Inconsistent > Consistent contrast across the whole brain (the two groups of participants were combined). However, no significant clusters survived the threshold.

fMRI Results: Multivariate analysis

To more closely analyse the neural responses to each condition, we ran further MVPA analysis. Since our hypothesis predicts that Inconsistent statements will be processed similar to Immoral statements or Negative statements, but significantly different to Consistent statements, we assessed the voxel-by-voxel correlation of activation between these conditions within the six associated ROIs.

Left Insula

We found a highly positive within-subject correlation across participants for Inconsistent-Immoral conditions ($t(34) = 23.98$, $p < 0.001$), Inconsistent-Negative conditions ($t(34) = 18.29$, $p < 0.001$), Inconsistent-Consistent conditions ($t(34) = 32.43$, $p < 0.001$), and Inconsistent-Control conditions ($t(34) = 18.82$, $p < 0.001$) (see Figure 2.5A, and also Supplementary (SM) Figure 2.S1A for heatmap of the average coefficients between all conditions).

To further investigate the relationship between correlations, we ran a series of planned corrected paired t-tests. It revealed a significant difference between Inconsistent-Consistent vs. the Inconsistent-Immoral ($t(34) = 5.59$, $p < 0.001$), Inconsistent-Consistent vs. Inconsistent-Negative ($t(34) = 4.08$, $p < 0.001$), and Inconsistent-Consistent vs. Inconsistent-Control ($t(34) = 4.53$, $p < 0.001$). There was no significant difference between the correlations of Inconsistent-

Immoral vs. Inconsistent-Negative (p = 1), or similarly Inconsistent-Immoral vs. Inconsistent-Control (p = 1). There was also no significant difference between the Inconsistent-Negative vs. Inconsistent-Control (p = 0.83; see Figure 2.5A). Interestingly, this demonstrates that the left insula may be more sensitive to political stimuli in general (both consistent and inconsistent) compared to immoral or negative stimuli.

To compare patterns of activity between groups (left wing versus right wing), we directly compared the coefficients between left and right wing subjects via independent t-test. No significant differences between any of the correlations were found (all ps > 0.05).

Right Insula

We found a highly positive within-subject correlation across participants for Inconsistent-Immoral conditions ($t(34) = 19.89$, $p < 0.001$), Inconsistent-Negative conditions ($t(34) = 12.66$, $p < 0.001$), Inconsistent-Consistent conditions ($t(34) = 19.34$, $p < 0.001$), and Inconsistent-Control conditions ($t(34) = 15.57$, $p < 0.001$) (see Figure 2.5B & SM Figure 2.S1B).

Planned paired t-tests revealed a significant difference between the average correlation coefficients for Inconsistent-Consistent vs. Inconsistent-Control ($t(34) = 3.21$, $p = 0.01$). No significant difference was found between the average correlation coefficients for Inconsistent-Immoral vs. Inconsistent-Negative (p = 1), Inconsistent-Immoral vs. Inconsistent-Consistent (p = 0.05), or Inconsistent-Immoral vs. Inconsistent-Control (p = 0.07) pattern of brain activation. Additionally, no difference between the Inconsistent-Negative vs. Inconsistent-Consistent (p = 0.06), or Inconsistent-Negative vs. Inconsistent-Control (p = 1) average correlation coefficients were found. This demonstrates that the right insula doesn't present different patterns of activation for processing politically inconsistent, consistent, immoral, or

negative material, but does present different patterns for general political material compared to neutral material (control).

To compare patterns of activity between groups (left wing versus right wing), we directly compared the transformed coefficients between left and right wing subjects via an independent t-test. No significant differences between any of the correlations were found (all ps > 0.05).

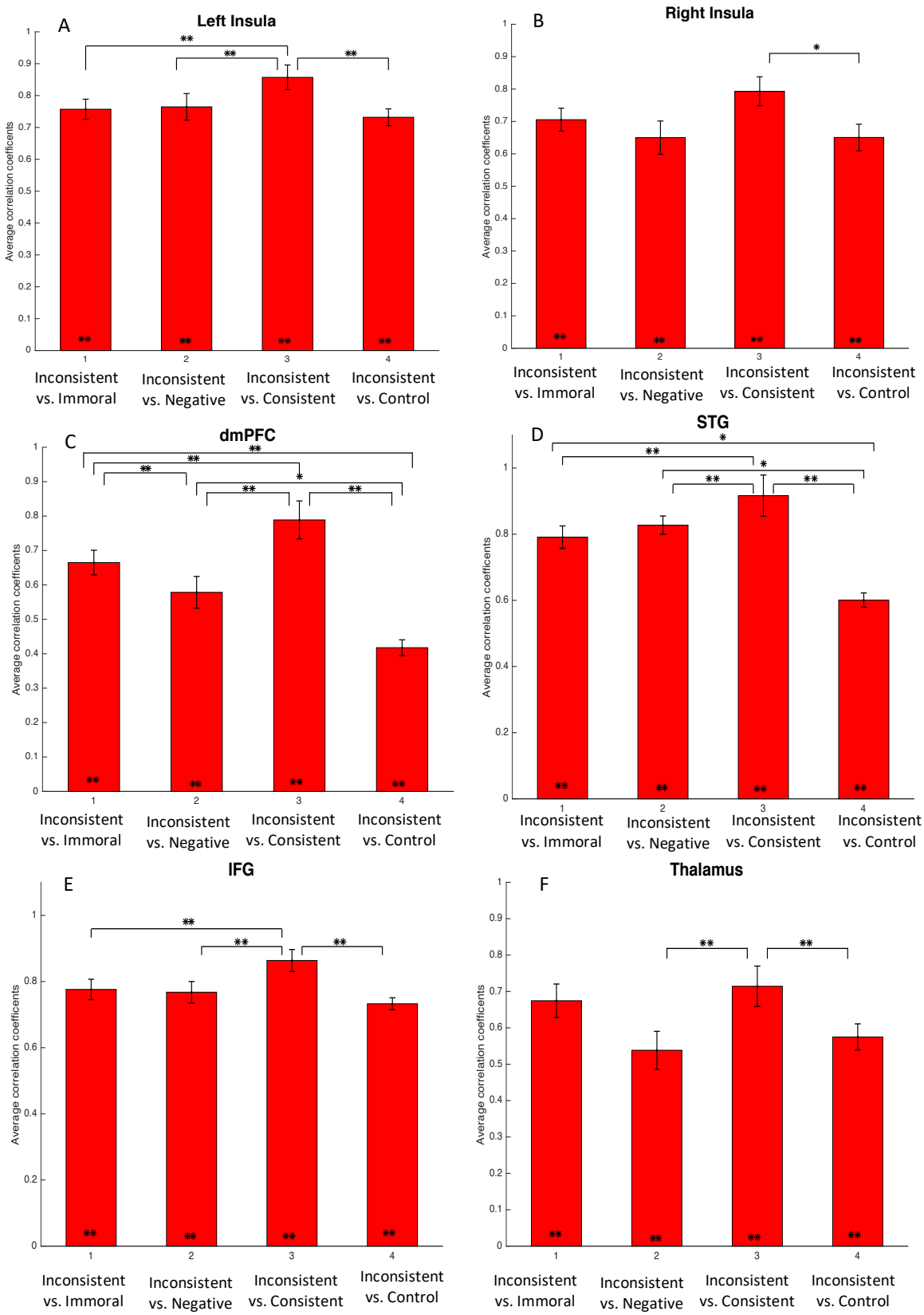


Figure 2.5. Bars represent average correlation coefficients in relevant ROIs. * $p < 0.05$, ** $p < 0.01$, below asterisks refers to one-sample t-test, above refers to paired t-tests. Error bars denote SEM.

dmPFC

We found a highly positive within-subject correlation across participants for Inconsistent-Immoral conditions ($t(34) = 18.56$, $p < 0.001$), Inconsistent-Negative conditions ($t(34) = 12.55$, $p < 0.001$), Inconsistent-Consistent conditions ($t(34) = 34.40$, $p < 0.001$), and Inconsistent-Control conditions ($t(34) = 7.57$, $p < 0.001$) (see Figure 2.5C & SM Figure 2.S1C).

Planned paired t-tests revealed a significant difference between the average coefficients for Inconsistent-Immoral vs. Inconsistent-Consistent ($t(34) = 4.89$, $p < 0.001$), and Inconsistent-Immoral vs. Inconsistent-Control ($t(34) = 4.75$, $p < 0.001$). There was also a significant difference between the average coefficients for Inconsistent-Negative vs. Inconsistent-Consistent ($t(34) = 6.50$, $p < 0.001$), Inconsistent-Negative vs. Inconsistent-Control $t(34) = 3.31$, $p = 0.01$), and Inconsistent-Consistent vs. Inconsistent-Control ($t(34) = 9.16$, $p < 0.001$). There was no significant difference between Inconsistent-Immoral vs. Inconsistent-Negative ($p = 0.13$). This demonstrates that the dmPFC may be more sensitive to political material in general, rather than immoral or negative material.

To compare patterns of activity between groups (left wing versus right wing), we directly compared the transformed coefficients between left and right wing subjects via an independent t-test. No significant differences between any of the correlations were found (all ps > 0.05).

STG

We found a highly positive within-subject correlation across participants for Inconsistent-Immoral conditions ($t(34) = 23.33$, $p < 0.001$), Inconsistent-Negative conditions ($t(34) = 29.88$, $p < 0.001$), Inconsistent-Consistent conditions ($t(34) = 42.42$, $p < 0.001$), and Inconsistent-Control conditions ($t(34) = 9.63$, $p < 0.001$) (see Figure 2.5D, SM Figure 2.S1D).

Planned paired t-tests revealed a significant difference between the average coefficients for Inconsistent-Immoral vs. Inconsistent-Consistent ($t(34) = 9.33$, $p < 0.001$), and Inconsistent-Immoral vs. Inconsistent-Control $t(34) = 2.93$, $p = 0.02$), as well as Inconsistent-Negative vs. Inconsistent-Consistent ($t(34) = 6.89$, $p < 0.001$), and Inconsistent-Negative vs. Inconsistent-Control ($t(34) = 4.97$, $p < 0.001$), and finally a significant difference between Inconsistent-Consistent vs. Inconsistent-Control ($t(34) = 8.20$, $p < 0.001$). There was no significant difference between the average coefficients for Inconsistent-Immoral vs. Inconsistent-Negative ($p = 0.09$; Figure 2.5D). This demonstrates the STG may be more sensitive to political material in general, rather than immoral or negative material.

To compare patterns of activity between groups (left wing versus right wing), we directly compared the transformed coefficients between left and right wing subjects via an independent t-test. No significant differences between any of the correlations were found (all ps > 0.05).

IFG

We found a highly positive within-subject correlation across participants for Inconsistent-Immoral conditions ($t(34) = 25.17$, $p < 0.001$), Inconsistent-Negative conditions ($t(34) = 23.69$, $p < 0.001$), Inconsistent-Consistent conditions ($t(34) = 47.50$, $p < 0.001$), and Inconsistent-Control conditions ($t(34) = 22.10$, $p < 0.001$) (see Figure 2.5E & SM Figure 2.S1E).

Planned paired t-tests revealed a significant difference between the average coefficients for Inconsistent-Immoral vs. Inconsistent-Consistent ($t(34) = 3.95$, $p < 0.001$), as well as Inconsistent-Negative vs. Inconsistent-Consistent ($t(34) = 4.20$, $p < 0.001$), and Inconsistent-Consistent vs. Inconsistent-Control $t(34) = 5.27$, $p < 0.001$). There was no significant difference between the average coefficients for Inconsistent-Immoral vs. Inconsistent-Negative ($p = 1$), and Inconsistent-Immoral vs. Inconsistent-Control ($p = 0.17$), or Inconsistent-Negative

vs. Inconsistent-Control (p = 0.16; Figure 6E). This demonstrates the IFG may be more sensitive to political material in general, rather than immoral or negative material.

To compare patterns of activity between groups (left wing versus right wing), we directly compared the transformed coefficients between left and right wing subjects via an independent t-test. No significant differences between any of the correlations were found (all ps > 0.05).

Thalamus

We found a highly positive within-subject correlation across participants for Inconsistent-Immoral conditions (t(34) = 14.68, p < 0.001), Inconsistent-Negative conditions (t(34) = 10.32, p < 0.001), Inconsistent-Consistent conditions (t(34) = 19.81, p < 0.001), and Inconsistent-Control conditions (t(34) = 10.35, p < 0.001) (see Figure 2.5F & SM Figure 2.S1F).

Planned paired t-tests revealed a significant difference between the average coefficients for Inconsistent-Negative vs. Inconsistent-Consistent (t(34) = 2.70, p < 0.001), and Inconsistent-Consistent vs. Inconsistent-Control (t(34) = 3.01, p = 0.01). There were no significant differences between the average correlation coefficients for Inconsistent-Immoral vs. Inconsistent-Consistent (p = 1), Inconsistent-Immoral vs. Inconsistent- Negative (p = 0.06), Inconsistent-Immoral vs. Inconsistent-Control (p = 0.08), or Inconsistent-Negative vs. Inconsistent-Control (p = 1; Figure 2.5F). This demonstrates that the thalamus processes general political material similar to immoral material, but differently to generally negative material.

To compare patterns of activity between groups (left wing versus right wing), we directly compared the transformed coefficients between left and right wing subjects via an independent t-test. No significant differences between any of the correlations were found (all ps > 0.05).

Brain-Behaviour Across Subject Correlations

In order to assess the relationship between our behavioural measures and any associated brain activity in greater depth, a series of brain-behaviour across subject correlations were conducted (*p* values are Bonferroni-Holm corrected for multiple comparison). We found no significant correlation between average political Intolerance scores versus brain activity for inconsistent compared to consistent trials (average beta values extracted from the Inconsistent > Consistent contrast) in all six ROIs (Figure 2.2; all *ps* > 0.16). Since political Intolerance was positively related to attitudinal strength, we additionally investigated the relationship with (left wing reversed) Orientation scores and brain activity, but found no significant correlation in the left insula ($p = 0.64$), right insula ($p = 0.64$), or IFG ($p = 0.24$). We did however find a significant positive relationship between Orientation scores and activity in the dmPFC ($r = 0.45$, $p = 0.04$; Figure 2.6A), STG ($r = 0.44$, $p = 0.03$; Figure 2.6B), and thalamus ($r = 0.45$, $p = 0.04$; Figure 2.6C). This demonstrates attitudinal strength is associated with increased activation for inconsistent versus consistent political material in these regions.

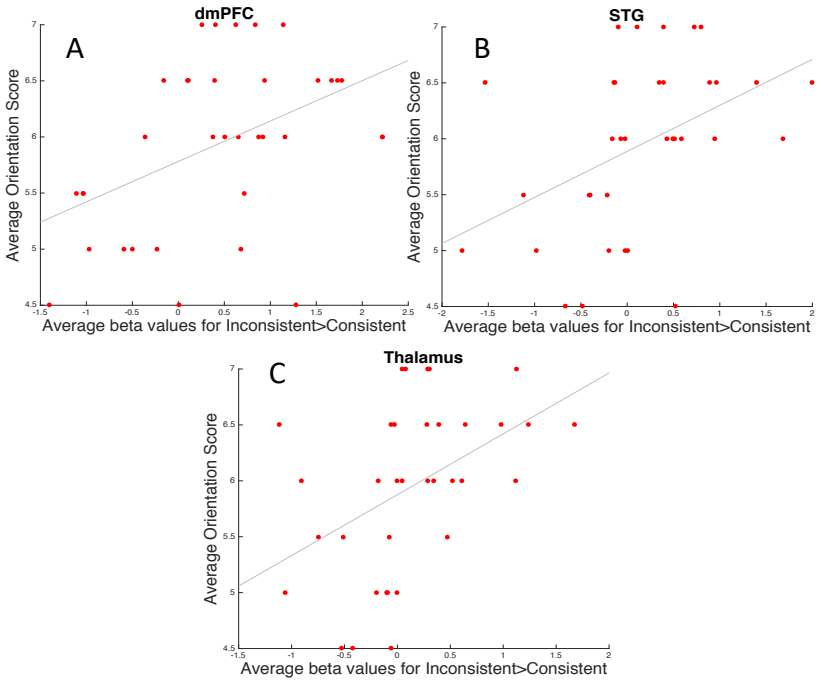


Figure 2.6. All panels demonstrate relationship between associated ROIs average activation for inconsistent compared to consistent political material versus average (left wing reversed) Orientation score.

DISCUSSION

The aim of this study was to examine any neural variation induced by political intolerance across left and right wing participants, whilst also investigating if material that is inconsistent with one's political orientation is processed similarly to immoral or negative material. This was examined by recruiting both left and right wing groups of politically engaged participants, and observing their neural responses to politically inconsistent stimuli in an fMRI scanner. Behaviourally, we found that attitude extremity was positively related to Intolerance scores across all participants indicating that stronger political attitude relates to the rejection of opposing political ideology's. We defined six ROIs (the bilateral insula, dmPFC, STG, IFG, and thalamus; see Figure 2.2) that are broadly related to emotions in general, and our univariate analyses first revealed that inconsistent political material elicited activation strength more similar to immoral rather than negative material in all of the six ROIs except the STG, suggesting inconsistent political material is perceived as morally disgusting. But, the correlation-based MVPA demonstrated that across ROIs, the pattern similarity was generally the highest between politically inconsistent vs. consistent material, indicating more similar processing for political material in general compared to apolitical. Furthermore, there was no clear difference between the inconsistent-immoral pattern similarity versus the inconsistent-negative pattern similarity, suggesting that politically inconsistent material was not particularly perceived as immoral. Importantly, we found no tangible difference in the processing of politically inconsistent information between our two political groups. This data, tentatively, supports the Ideological Conflict Hypothesis, the notion that intolerance derives from opposing ideology, not specific characteristics of political orientation.

All ROIs excluding the right insula demonstrated significantly increased activation for politically inconsistent material compared to neutral material across participants. These results fall in line with previous data demonstrating the role of the anterior insula and mPFC in

processing politically opposing material (Kaplan et al., 2007, 2016; Knutson et al., 2006; Westen et al., 2006). Activation of the thalamus for politically inconsistent material also supports preceding work, with the role of the thalamus mainly being described as an information transmission hub, relaying critical information from external (top down, environmental) and internal (bottom up) cues (Saalmann & Kastner, 2011).

Further, the dmPFC, STG, and thalamus show stronger activation for inconsistent material the more extreme the participants attitude was, also echoed in previous literature such as Kaplan et al., (2007) who found neural responses to candiates faces varied in regard to feelings towards candidates, and Knutson et al., (2006) who found feelings towards candidates used in an IAT were related to increased frontopolar activity. The dmPFC can generally be attributed to behaviour/action monitoring and selection, and the social evaluations of others (Rushworth, Buckley, Behrens, Walton, & Bannerman, 2007; Rushworth, Walton, Kennerley, & Bannerman, 2004; Talati & Hirsch, 2005), particularly left lateralisation (Talati & Hirsch, 2005). More, the STG aside from being a key component in language processing (Bigler et al., 2007), is also implicated in regulating social cognition via behavioural monitoring and assessment (Adolphs, 2003; Bigler et al., 2007; Takahashi et al., 2004). Overall these regions, including the thalamus mentioned previously, seem to represent the transmission and processing of information relevant to assessing social items and subsequent behaviour. Thus, as individuals with stronger attitudes (regardless of political orientation) showed stronger activity within these regions for inconsistent compared to consistent political material (Figure 2.6), this suggests a more prominent predictor of intolerance, and tool for future work wishing to examine the neural correlates of intolerance, is attitude extremity, rather than specific attitudinal orientation.

Interestingly however, our experiment finds little evidence to suggest our ROIs process politically inconsistent material differently to consistent. Only the dmPFC showed increased

activation for politically inconsistent material compared to consistent material, with further multivariate analysis showing the pattern of activation to be similar for general political material (consistent and inconsistent). Thus, it seems that politically inconsistent information isn't processed particularly similar to immoral *or* negative material, but may be handled via mechanisms more specific to political stimuli in general.

Considered in conjunction with this is the fact that political material in general may have been more arousing to participants overall. Due to participants being especially politically engaged, and the advert of the study itself being centred around politics, it's quite likely the participants were generally more interested in/anticipated more the political trials. This is reinforced with research by Cunningham, Raye, and Johnson (2004) who demonstrate the involvement of the insula regarding both negative *and* positive attitude valence, signifying the relevance for intensely valent material in general, not just negative/conflicting material. However, though this supports univariate findings regarding activation strength, there is research that infers distinct activation patterns regarding basic emotion (Vytal & Hamann, 2010), and so this interpretation may not necessarily be applied to our multivariate finding, as it remains somewhat unclear whether socially consistent versus inconsistent (i.e. positively or negatively valent information) is encoded similar, and thus elicits similar activation patterns, across our ROIs.

Since all ROIs examined demonstrated a significant positive correlation between conditions via MVPA (indicating the pattern of processing to be similar across all conditions), it seems relevant to consider more general processes participants undertook, for example sentence comprehension. A recent review (from 37 studies) indicates the engagement of the left inferior frontal and posterior temporal regions, and right insula, for the comprehension of complex syntax (Walenski, Europa, Caplan, & Thompson, 2019). Due to the complex nature of our stimuli, it is likely multiple regions will co-ordinate in a similar pattern in order to

process the basic information presented, which is why the use of control stimuli is essential when assessing complex (particularly social) neural processes. Furthermore, all trials undertook by participants contained images of faces, and research has shown viewing faces can stimulate right insula activity (Kircher et al., 2000), aptly where the least variation in activation pattern across conditions is seen. Additionally, the STG and IFG are also implicated in the input of the facial responsive network (for example see Haxby, Hoffman, & Gobbini, 2000), and impression formation for human faces compared to objects is implicated within neural networks in the mPFC area including the STG (Mitchell, Neil Macrae, & Banaji, 2005).

A key finding from this study is that no tangible difference is seen in the neural correlates of left versus right wing subjects when processing inconsistent political material, despite a substantial section of previous literature that might allude to such. This cautiously supports the Ideological Conflict hypothesis, but emphasis should be placed on the small, imbalanced sample size that reduced the current experiments statistical power. Of the behavioural differences that are present, this actually indicates increased political intolerance from our left wing sample, but there are several important factors to take into account. Firstly, the (left wing reversed) Orientation scores (a basic measure of attitude strength) for left wing participants were significantly more extreme than for right wing participants. Research has demonstrated that more extreme attitudes tend to induce higher levels of political Intolerance (Alter, Oppenheimer, & Zemla, 2010; Fernbach et al., 2013; van Prooijen & Krouwel, 2017). This is supplemented by our findings that demonstrate a positive relationship between attitude extremity and i) political Intolerance, and ii) higher average activation for politically inconsistent compared to consistent material in the dmPFC, STG and thalamus.

Although no convincing neural or behavioural differences regarding the intolerance towards opposing political ideas were found between our groups, indicating processes perceiving inconsistent political material may be more similar, our study does still provide

some behavioural evidence demonstrating classic differences between political left versus right general characteristics. Mainly, our left wing participants provided higher Social Orientation scores relative to Economic, and our right wing participants provided higher Economic Orientation scores relative to Social. This suggests an asymmetry in the focus on social versus economic issues between our groups. This compliments previous work outlining the fundamental traits of political liberalism and conservatism. For example, the left is characterised by high levels of openness to experience, novelty seeking (Jost et al., 2009, 2003), and higher cognitive flexibility/ability (Adorno et al., 1950; Crisp & Meleady, 2012; Jost et al., 2003). These traits compliment an increased focus on social values in our left wing participants. Similarly, the right is characterised by high levels of sensitivity to threat (Oxley et al., 2008), traditionalism (Jost et al., 2003), a strong desire for order (Carney, Jost, Gosling, & Potter, 2008), and increased organisation skills (Caprara, Schwartz, Capanna, Vecchione, & Barbaranelli, 2006), complimenting an increased focus on more economic and structural values.

It should be considered an important factor that the participants for this experiment were all students, a demographic not utilised by Crawford and Pilanski (2014) who found no difference in intolerance between liberals and conservatives. Universities are notoriously left wing environments. For example, one survey suggested that eight out of ten university lecturers in Britain identify as left wing (Turner, 2018). This could mean the image of right wing views are stigmatised in a university environment, having a detrimental effect on right wing participants who may feel more hesitant about expressing their opinions. Conversely, it should also be noted that Wetherell et al., (2013) did utilise a student population for their first of two experiments, and didn't find a significant difference between liberals and conservatives levels of political intolerance. Hence, what may additionally be important to consider was the current political climate in 2016 (encapsulating some mass right wing populist movements; i.e.

"Brexit", the US election of President Trump). Arguably in this context, right wing positions were *more* unpopular in liberal-orientated contexts. An example of this is the reported strong opposition to Britain's leaving the EU amongst British Universities, with university graduates reported to be the most likely demographic to vote remain (see Kirk & Dunford, 2017).

Overall, these results seem promising in potentially supporting the Ideological Conflict hypothesis (Brandt et al., 2014), the notion that intolerance derives from opposing world view rather than specific right-wing orientated traits. But, as mentioned previously, research should seek to validate findings utilising a larger sample with more power, and attitudinally matched groups. As this effect is mediated in general by attitudinal strength, attitude extremity may be a more accurate predictor to isolate specific neural and cognitive processes involved in the intolerance to opposing political material. Future research wishing to further examine any variation in the predictors of Intolerance across the left/right divide should utilise where possible groups of matched attitudinal strength.

CONCLUSION

In summary, the key findings from this study are as follows: i) no neural variation in the way left and right wing individuals process inconsistent political material is observed, tentatively suggesting there is no difference in how the two groups process material that is inconsistent with their political view, ii) attitude extremity rather than attitude orientation may be a better predictor of intolerance to further isolate specific neural mechanisms, iii) the left insula, dmPFC, STG, and IFG exhibit neural correlates more similar for political material in general compared to apolitical material, suggesting more exclusive mechanisms for the processing of socio-political information. Together our results provide more basis into not only the neural variation between political groups previously not directly measured, but also add support to the specific neural processes relevant to processing political material in general.

Chapter 3

The neural response to conflict

Since the previous chapter worked to identify some of the neural principles involved in processing politically inconsistent information, the next chapter of my doctoral research aims to focus on the fundamental neural principles associated with processing generally conflicting information. Conflicting information in this instance involves anything that is incompatible or not in keeping with one's current predisposition, beliefs, or expectations.

Classic examples demonstrating the general effect conflicting information has on our ability to process information come from experiments utilising a Stroop task (Stroop, 1935), in which the ACC tends to be specifically implicated upon response conflict (when a word colour name doesn't match the ink colour) (Barch et al., 2001; Bench et al., 1991; Fan et al., 2002; Kerns et al., 2004; Lee et al., 2004; Leung et al., 2000). This subsequently lead to the development of a conflict monitoring model (Botvinick, Braver, Barch, Carter, & Cohen, 2001). This model postulates the dACC continuously analyses current information for possible response conflict arising from interference between separate processing streams. The model then proposes a cognitive control system comes into play once conflict is detected, routed in the dlPFC, by biasing information processing mechanisms in relevant posterior brain regions.

fMRI studies also demonstrate using social conformity tasks that the pMFC (ACC and dmPFC particularly) tracks the conflict ensued by the difference between an individual's versus wider group's opinion, and the subsequent shift of opinion towards the wider group (Campbell-Meiklejohn et al., 2010; Izuma & Adolphs, 2013; Klucharev, Hytönen, Rijpkema, Smidts, & Fernández, 2009; Wu et al., 2016). Interestingly, using the Multi-Source Interference Task (a Stroop-like task where participants must quickly identify particular cues whilst systematic interference takes place via several additional cues) (MSIT; Bush & Shin, 2006), Izuma and Adolph (2013) also demonstrate a region within the pMFC, the pre supplementary motor area

(pre SMA) was specifically active for general response conflict, but wasn't for socially desirable versus undesirable outcomes, which was associated with the dmPFC (a more anterior region of the pMFC). This indicates the social conflict of desirable outcomes and reality may elicit distinct neural responses.

Further examples of more socially pertinent paradigms include fMRI investigations into moral conflicts, assessing the neural correlates when faced with acting for self versus collective interest. Here it's found the ACC, prefrontal cortex, parietal lobe, and temporoparietal junction are more active when faced with morally conflicting trials (Emonds, Declerck, Boone, Vandervliet, & Parizel, 2012), and Greene, Nystrom, Engell, Darley, and Cohen, (2004) found increased activity in the ACC and dlPFC for difficult compared to easy moral conflicts.

As can be inferred from the above, in the instance of social conformity participants tend to resolve conflict (i.e., difference between one's and group's opinions) by changing their behaviour (i.e. shifting opinions or preference to that of the wider group). What remains unclear still are the more specific mechanisms in the pMFC regarding the processing of social conflict. Primarily, if the pMFC is involved in processing conflict alone (conflict detection), or also the impeding adjustment of behaviour (conflict resolution).

Chapter 4

Elucidating the role of the posterior medial frontal cortex in social conflict processing

Stephanie J. Wake[1], Ryuta Aoki[3], Kiyoshi Nakahara[4] & Keise Izuma[1,2,]

[1] Department of Psychology, University of York, Heslington, York, YO10 5DD, UK.

[2] Department of Psychology, University of Southampton, University Road, Southampton, SO17 1BJ, UK

[3] Research Institute for Future Design, Kochi University of Technology, Eikokuji, Kochi 780-8515, Japan

[4] Research Center for Brain Communication, Kochi University of Technology, Kami, Kochi 782-8502, Japan.

ABSTRACT

A fundamental function of the brain is learning via new information. Studies investigating the neural basis of information-based learning processes indicate an important role played by the posterior medial frontal cortex (pMFC) in representing conflict between an individual's expectation and new information. However, specific function of the pMFC in this process remains relatively indistinct. Particularly, it's unclear whether the pMFC plays a role in the detection of *conflict* of incoming information, or the *update* of their belief after new information is provided. In an fMRI scanner, twenty-eight Japanese students viewed scenarios depicting various pro-social/anti-social behaviours. Participants rated how likely Japanese and South Korean students would perform each behaviour, followed by feedback of the actual likelihood. They were then asked to rerate the scenarios after the fMRI session. Participants updated their second estimates based on feedback, with estimate changes more pronounced for favourable feedback (when the interaction between scenario type and feedback paints the individual in a more favourable light i.e. higher likelihood of pro-social behaviour than expected) despite nationality, indicating participants were willing to view other people favourably. The fMRI results demonstrated activity in a part of the pMFC, the dorsomedial prefrontal cortex (dmPFC), was correlated with social conflict (difference between participant's estimate and actual likelihood), but not the corresponding belief update. Importantly, activity in a different part within the dmPFC was more sensitive to *unfavourable* trials compared to favourable trials. These results indicate sensitivity in the pMFC (at least within the dmPFC) relates to conflict between desirable outcomes versus reality, as opposed to the associated update of belief.

Key Words: pMFC, social attitudes, favourability bias, conformity, learning

INTRODUCTION

Procuring knowledge via new information is one of the most important functions of the brain. We update our beliefs, knowledge and/or attitudes based on semantic factual information (e.g., how likely you are to become ill) as well as what other people think (i.e., social conformity). A number of past neuroimaging studies have investigated the neural mechanisms behind information-based learning processes, and currently available evidence converge to indicate an important role played by the posterior part of the medial frontal cortex (pMFC), particularly the dorsomedial prefrontal cortex (dmPFC) and dorsal anterior cingulate cortex (dACC), in representing the conflict between an individual's expectation and new information.

The pMFC is known to play a key role in processing reward prediction error (i.e., the difference between actual and predicted reward) in reinforcement learning tasks (specifically the ACC) (Sambrook & Goslin, 2015), and a number of neuroimaging studies have indicated that the pMFC plays a wider role, being involved in information-based learning in a variety of both social and non-social settings where there is no reward. For example, using a social conformity task, a functional magnetic resonance imaging (fMRI) study Klucharev, Hytönen, Rijpkema, Smidts, and Fernández (2009) demonstrated that the rostral cingulate zone, a part of the ACC, tracked the discrepancy between individual's versus group's opinion so that the larger the conflict between one's and group's opinions, the higher the activity. This result has been replicated by other fMRI studies (Campbell-Meiklejohn et al., 2010; Izuma & Adolphs, 2013; Wu et al., 2016). Similarly, a number of electroencephalography (EEG) studies on social conformity (Chen et al., 2012; Huang et al., 2014; Kim et al., 2012; Schnuerch et al., 2014; Schnuerch & Gibbons, 2015; Shestakova et al., 2012) observed electrophysiological responses over the pMFC that track the conflict between one's versus group's opinion. The electrophysiological responses resemble the feedback-related negativity (FRN) signal, which is related to reward prediction error and is considered to be generated in the ACC (Holroyd &

Coles, 2002; Sambrook & Goslin, 2015). Furthermore, more recently, Pine et al., (2018) demonstrated that the dmPFC, is involved in prediction error in learning based on semantic factual information.

Izuma and Adolphs (2013) further demonstrated that the pMFC doesn't simply represent the conflict between one's and others' opinion, but rather, it represents the conflict posed from desired versus undesired outcome (Izuma, 2013). Izuma and Adolphs (2013) first replicated Klucharev et al.'s (2009) findings showing the pMFC (specifically the dmPFC) tracked the conflict between participant's and their fellow students' (participant's "liked" group) opinions. However, this pattern was completely reversed if it was an opinion of a "disliked" group; the pMFC activity was higher when their opinion was more *similar* to sex offenders' (disliked group) opinion. Thus, the results suggest the pMFC doesn't solely represent the distance between one's and others' opinion, but more embodies the divergence from desirable outcomes.

Although a number of studies have demonstrated that pMFC activity reflects the discrepancy between an individual's expectation (or opinion) and new information (or more broadly, the discrepancy between a desirable or ideal outcome, and reality), the exact roles of the pMFC in information-based learning still remains to be fully elucidated. More specifically, it remains unclear whether the pMFC plays a specific role in the detection of *conflict* of incoming information (with the dACC particularly involved in conflict monitoring and successive cognitive control; Mansouri et al., 2017; Shenhav et al., 2013), or is associated with the *update* of their belief after new information is provided. In previous studies, these two processes often co-occurred- making it difficult to disentangle them. For example, in a typical social conformity study, the larger the conflict between one's versus group's opinions, the more an individual conforms to the group's opinion (i.e., the greater update of their opinion).

Accordingly, the current study aimed to shed a new light on the role of the pMFC by utilising cognitive bias, extending the findings of Izuma and Adolphs (2013). Numerous studies

in psychology have demonstrated that we don't process information objectively, rather how we process new information is heavily affected by various cognitive biases. For example, as a general rule we tend to seek and formulate our attitudes based on information that already aligns with our own ideals, a phenomenon known as confirmation bias (Knobloch-Westerwick et al., 2015; Lord et al., 1979; Sunstein et al., 2016). Thus, how we update our belief depends on whether new information is consistent with how an individual already sees the world. Appropriately, by utilising a cognitive bias, we can dissociate the level of conflict from the level of belief updating (e.g., the same degree of conflict can predict different levels of belief updating dependent on whether it is consistent with their pre-existing ideals).

Confirmation bias here was elicited using an intergroup paradigm, specifically Japanese participants perceptions of other Japanese individuals (in-group) versus South Korean individuals (out-group), whom historically have a tense relationship (see Izuma et al., 2019; Lee, 1985). The vast social body of research regarding intergroup relations informs us that general favouritism towards the in-group and derogation towards an out-group tends to be a common nature of human group behaviour (for example Tajfel, 1982; Tajfel, 2010). Extensions to neuroscience research have been made increasingly apparent (for a recent review see Molenberghs & Louis, 2018; Hackel et al., 2017). A recent example comes from Lin et al., (2018), who found that after participants rated emotional stimuli in the scanner, they were more likely to change their evaluations to be more similar to the evaluations other in-group members made compared to the out-group. This shift was tracked by neural activity in the ventral striatum, dmPFC, mPFC, posterior superior temporal sulcus (pSTS), temporal pole, amygdala and insula (see also Huang et al. 2019). Thus, we applied an intergroup context to promote confirmation bias, directly manipulating the level of bias participants are presented with.

In the study, Japanese university students viewed a series of scenarios which describe either a pro-social or anti-social behaviour inside an MRI scanner. Their task was to estimate

how typical Japanese and South Korean students answered a series of questions relating to how they would respond in said scenarios (Figure 1). After they gave their rating, participants were presented with the rating given by Japanese or South Korean students (i.e., what percentage of Japanese or South Korean students were willing to perform the pro- or anti-social behaviour). After participants had gone through all scenarios and feedback, they were then asked to rerate the scenarios as an experimental task outside of the scanner to index the level of belief updating.

Behaviourally, we expected that how much individuals updated their belief about Japanese and South Korean students depends on their attitudes toward Japan and South Korea, respectively, and the pro-social nature of the feedback presented. To the extent that our Japanese participants have positive attitudes toward Japan, they would update their belief about Japanese students more if new information allows them to see other Japanese students more favourably (e.g., if more Japanese students were willing to perform a pro-social behaviour than expected). We expected a similar pattern for the South Korea condition, but this favourability bias would be less pronounced because of participants' less positive attitudes toward South Korea (out-group) compared to Japan (in-group) (i.e., participants' would be more willing to view in-group members favourably compared to out-group members).

Furthermore, the study aimed to test the two competing hypotheses regarding pMFC activity, specifically the dmPFC. First, if the dmPFC encodes the conflict between a desirable state versus reality, its activity should be more sensitive to the difference between one's estimate and actual feedback when the feedback is in an *unfavourable* direction (conflict hypothesis). In contrast, if the pMFC plays a role in belief updating, its activity should be more sensitive to the difference when the feedback is in a *favourable* direction where we expect a larger update of their belief (update hypothesis).

METHOD

Participants

Twenty-nine right-handed Japanese students with no psychiatric history were recruited via a participant pool at the Kochi University of Technology. One participant was excluded from the analysis due to excessive head motion (i.e., >3mm). The final sample consists of 28 participants (male = 16, female = 12; mean age = 20.3). Note that due to a technical fault with the scanner, for one subject, fMRI data after 6 minutes of the first session were not obtained. Accordingly, for the first session of this subject, the fMRI data analysis included 144 images (it should have been 214 images). In this session, the subject still continued the task without being scanned for approximately 3 minutes so that our behavioural data analysis included all trials. All participants gave written informed consent for participation, and ethics approval for the study was granted by the Kochi University of Technology Ethics Board.

Procedure & Task

Participants were told they would view a series of scenarios which describe either a pro-social or anti-social behaviour (e.g. "*Japanese students from University F were presented with the scenario of seeing racist material towards South Korean people on social media, and asked if they condoned this*", for full list of scenarios used see Supplementary Materials, Appendix 2) inside an fMRI scanner, and it was their task to estimate how typical Japanese and South Korean students answered a series of questions relating to how they would respond in said scenarios. They were asked to rate on a scale of 0%-100% in increments of 5 using a button box with three buttons. They used the index finger to increase the rating by 5%, the middle finger to reduce it by 5%, and the ring finger to give a final decision. All participants used their right hand to give responses. After they gave their rating, participants were presented with the "actual" rating given by Japanese or South Korean students, hereby referred to as *feedback* (see

Figure 4.1 for visual of a complete trial). Although participants were led to believe that the feedback was real, in reality it was determined by a simple algorithm. Participants were exposed to 4 types of scenarios (2 [pro- versus anti- social] × 2 [Japan versus South Korea]), with a feedback trial that was higher or lower than the participant's first estimate. Our algorithm, computed via Matlab, ensured that feedback created roughly equal numbers of conditions across sessions, with a possible difference between participants' first ratings and feedback ranging from 5 to 30. The fMRI session consisted of a total of four runs, each consisting of 28 experimental trials plus 1 catch trial (where we presented feedback that coincided with participant's first estimates). Participants were presented with the initial scenario for 3 seconds, with no limit when providing their ratings on how likely the group in question would partake in such scenario. Subjects response was highlighted for 1 second before feedback was presented for 2 seconds.

A total of 56 scenarios (plus 4 catch trials) were used in the fMRI experiment, and these scenarios were selected by a pilot study with an independent sample of n = 17 (mean age = 20.2, 9 males) from the Kochi University of Technology. In the pilot study, participants were asked to rate how likely a group of Japanese and South Korean students would respond to a total of 112 (56 Japanese and 56 South Korean) scenarios, as well as rate how positive/negative (valence rating) and relevant each scenario was on a scale of 1-7. Scenarios that presented extreme (ratings that fell outside of the bottom 7% and top 90%) ratings (how likely the target group in question responded) were discarded so as to reduce the effect of participants inevitably providing less extreme ratings in a subsequent second rating task, known as the regression-to-the-mean effect (RTM) which continually illustrates when repeated measures designs are used extreme values at the first measurement tend to approach the mean at the succeeding measurement (Galton, 1886; Yu & Chen., 2015). Scenarios were additionally matched for valence and relevance. This data was also used to generate extra scenarios that resembled and

replicated the general theme of accepted scenarios, yielding a total of 28 positive Japanese scenarios, 28 negative Japanese scenarios, 28 positive South Korean scenarios, and 28 negative South Korean scenarios. Note that participants view the same positive scenarios for both the Japan and South Korea conditions, likewise for negative scenarios (i.e., "*Japanese students from University F were presented with the scenario of seeing racist material towards South Korea…"* versus "*South Korean students from University C were presented with the scenario of seeing racist material towards Japan…"* - the only aspect manipulated is the nationality of the students depicted in the scenario).

After the main fMRI session, participants were asked to re-rate all 112 scenarios they viewed in the scanner. This was to assess the effect of learning or update. In addition, they rated each of the 56 scenarios using a 7-point scale on how socially desirable the behaviour depicted in each scenario was, excluding any nationality information (that of previous students completing the task and also the person depicted in the scenario) (1 = extremely socially undesirable, 4 = neither socially desirable nor undesirable, 7 = extremely socially desirable).

To assess their implicit attitudes toward Japan and South Korea, participants were asked to complete an Implicit Association Test (IAT) (Greenwald, McGhee, & Schwartz, 1998). The IAT included eight positive (e.g., Joy, Love, Wonderful) and eight negative words (e.g., Agony, Terrible, Nasty), all words were translated into Japanese. The Japan category included typical Japanese names (e.g., Shima, Nakata, Ono) whilst the South Korean category included typical Korean names (e.g., Han, Kim, Myong). All Japanese and South Korean names were matched on word length. Finally, their explicit attitudes toward Japan and South Korea were measured using a semantic differential scale. Participants rated Japan and South Korea on six bipolar dimensions using a 7-point scale; ugly-beautiful, bad-good, unpleasant-pleasant, honest-dishonest, foolish-wise, awful-nice and unfavourable-favourable. Finally, after completing a demographics questionnaire, to help ensure our experimental stimuli was

efficient, participants were asked if they doubted anything during the experiment. They were debriefed, thanked and paid 2,000 yen for their participation.

Figure 4.1. (**A**) Example of a complete South Korean trial (scenario, question/first rating, feedback) utilised for fMRI stimuli, as seen by participants inside the scanner. Each trial started with a scenario presentation (description of a pro- or anti-social behaviour) for 3 seconds, after which participants' were asked to give their first estimation of how likely the person in question

(Japanese vs. South Korean student) rated they would partake in said behaviour (in which they had no time limit). After, the estimate was highlighted in yellow for 1 second followed by feedback presentation (the "true value") for 2 seconds. (**B**). Visual representation of Absolute Gap and Update scores. (**C**). Example of 4 scenario types depicted via a pro-social scenario. Feedback was reversed in order to create the same conditions for anti-social scenarios.

fMRI Data Acquisition

All fMRI data was acquired using a Siemens 3.0 Tesla Verio scanner with a 32 channel phased array head coil. For functional imaging, interleaved T2*- weighted gradient-echo echo-planar imaging (EPI) sequences were used to produce 40 contiguous 3mm thick trans-axial slices covering nearly the entire cerebrum (repetition time [TR] = 2,500ms; echo time [TE] = 25ms; flip angle [FA] = 90°; field of view [FOV] = 192 mm; 64 × 64 matrix; voxel dimensions = 3.0 × 3.0 × 3.0mm). A high-resolution anatomical T1-weighted image (1 mm isotropic resolution) was also acquired for each participant.

fMRI Data Pre-processing

The fMRI data was analysed using SPM12 (Welcome Department of Imaging Neuroscience) implemented in Matlab (Math Works). Before data processing and statistical analysis, we discarded the first four volumes to allow for T1 equilibration. Head motion was corrected using the realignment program of SPM12. Following realignment, the volumes were normalised to MNI space using a transformation matrix obtained from the normalisation of the first EPI image of each individual participant to the EPI template using an affine transformation (resliced to a voxel size of 2.0 × 2.0 × 2.0mm). The normalised fMRI data were spatially smoothed with an isotropic Gaussian kernel of 8 mm (full-width at half-maximum).

fMRI Data Analysis

We used two general linear models (GLM) to analyse the fMRI data; one GLM was intended to identify brain regions correlated with the absolute differences between participant's

estimate and feedback (hereby referred to as: *Absolute Gap*, see Figure 4.1B), and the other GLM was to explore brain regions correlated with the behavioural *Update* (difference between the first estimate and the second estimate, see Figure 4.1B).

We used a parametric modulation analysis to investigate the relationship between trial-by-trial Absolute Gap scores and regional brain activity. We analysed the fMRI data based on a 2 (Japan or South Korea) × 2 (favourable or unfavourable) design, yielding the four following conditions: 1) Japan-Favourable, 2) Japan-Unfavourable, 3) South Korea-Favourable, and 4) South Korea-Unfavourable, and data was first divided into four sets accordingly. The factor of favourable-unfavourable refers to the interaction between the valence of presented scenarios (positive or negative) and the feedback given in relation to participants first estimates (if this was better or worse than participants initial expectations), and whether this combination comes across as overall pro-social or anti-social. For example, a favourable trial would be depicted by higher feedback in a positive scenario (i.e., Japanese or South Korean students are *more* willing to act pro-socially than participants expected) or lower feedback in a negative scenario (i.e., Japanese or South Korean students are *less* willing to act anti-socially than participants expected). Accordingly, the first model included: 1) each trial presentation (duration = total time from onset of initial scenario presentation to onset of feedback presentation), 2) Feedback presentation in Japanese favourable trials (duration = 2 sec), 3) Feedback presentation in Japanese favourable trials modulated by Absolute Gap, 4) Feedback presentation in Japanese unfavourable trials (duration = 2 sec), 5) Feedback presentation in Japanese unfavourable trials modulated by Absolute Gap, 6) Feedback presentation in South Korean favourable trials (duration = 2 sec), 7) Feedback presentation in South Korean favourable trials modulated by Absolute Gap, 8) Feedback presentation in South Korean unfavourable trials (duration = 2 sec), 9) Feedback presentation in South Korean unfavourable trials modulated by Absolute Gap, 10) Catch trial presentation (regressor of no

interest) (duration = total time of catch trial from initial scenario presentation onset to the end of feedback presentation). This analysis yielded the four main contrast images (all conditions modulated by Absolute Gap) used for second level analysis. Other regressors that were of no interest, such as six motion parameters, the session effect, and high-pass filtering (128 sec) were also included.

The second GLM is similar to the first except we used the behavioural *Update* (the difference between the first vs. second estimates) as opposed to *Absolute Gap* (the difference between the first estimate vs. feedback) as a parametric regressor. Because the simple difference between the two estimates is susceptible to the RTM effect (Izuma & Adolphs, 2013; Yu & Chen, 2015), in order to remove the change between the first vs. second estimates which can be explained by the RTM effect, we first ran a linear regression analysis within each participant to estimate the RTM effect for each participant. The regression model used all 112 trials and included participant's first estimates as the only predictor variable, and Update as the dependent variable. All participants showed a negative beta value for first estimates (e.g., the higher the first estimate, the more likely participants decrease their estimate on the second rating task), and at group level, it was significantly negative ($t(27) = -11.92$, $p < 0.001$), indicating the existence of the RTM effect. Within each participant, for each trial, we computed the Update scores predicted by the RTM effect and subtracted it from the actual Update scores (actual Update scores - Update scores predicted by the RTM effect). We then used the new controlled Update scores as parametric modulators in the second GLM. The same set up was utilised yielding the same contrast images to be used for second level analysis. For all fMRI analysis, a whole-brain statistical threshold was set at $p < 0.001$ voxel wise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple comparisons).

In addition to these two main GLMs, we also ran three additional GLMs (see Supplementary Materials, Appendix 2, for the full details and results of these GLMs); one

addressed the effect of the "general favourability" of feedback (i.e. if feedback indicated more people are willing to engage in a socially desirable behaviour or less people are willing to engage in an anti-social behaviour, *regardless of participant's expectations*). The second GLM incorporated both Absolute Gap and Update in a single GLM, and the third incorporated Update and Favourability in a single GLM to assess the interaction of Update x Favourability on brain activity.

Behavioural Data Analysis

For the IAT, a score for each participant was calculated using the D-score algorithm developed by Greenwald, Nosek, and Banaji (2003). Positive IAT D-scores indicate more positive implicit evaluation of Japan relative to South Korea. Semantic differential scores for each participant were computed by averaging the six bipolar scales separately for Japan and South Korea.

To calculate the effect of feedback on the extent participants updated their second estimates, two multiple regressions (one for Japanese trials, and one for South Korean trials) were run to analyse behavioural data. Both included predictor variables: 1) First Estimates, 2) Gap (feedback - first estimate, not absolute value), 3) Favourability (dummy coded as favourable = 1 and unfavourable = 0), and 4) Gap × Favourability. All predictors were centred by subtracting the mean value from each score to evade multicollinearity. The dependent variable was Update (second estimate – first estimate).

We additionally ran a similar analysis to assess the effect of "general favourability" of trials as mentioned above (see Supplementary Materials for the full details and results of this analysis).

Due to our stimuli incorporating scenarios that do versus don't involve the in-group in some form (i.e. "*... If you saw racist material towards Japanese people on social media, would*

you feel positive about it? ", versus, "*... Do you believe it is acceptable that when intoxicated at a party people sometimes vandalise property?* "), we conducted analysis to compare any potential confounds from this. We divided the data into scenarios that did involve the in-group (n=15), and scenarios that didn't (n=13). The same analysis as described above for both Japanese and South Korean trials was applied within each set of data, for full details, see Supplementary Materials.

RESULTS

Behavioural Results

Attitudes towards Japan versus South Korea

We first found that, not surprisingly, Japanese participants' explicit evaluations of Japan were significantly more positive than those of South Korea: ($t(27) = 7.95$, $p < 0.001$, Cohen's d = 1.97) (Figure 4.2A). We further demonstrate that explicit evaluations of Japan are significantly positive (by examining how different the mean score was from the midpoint of the scale: [$t(27) = 11.55$, $p < 0.001$]), and that those of South Korean were significantly negative ($t(27) = -2.11$, $p = 0.04$). Additionally, IAT scores were significantly positive ($t(27) = 4.14$, $p < 0.001$, Cohen's d = 0.80) (Figure 4.2B), indicating more positive implicit evaluations of Japan relative to South Korea. No significant correlation was observed for implicit evaluations and explicit evaluations (Japanese minus South Korean mean scores) ($r = 0.10$, $p = 0.62$), and no significant correlation was observed for explicit evaluations between Japan versus South Korea ($r = 0.16$, $p = 0.41$).

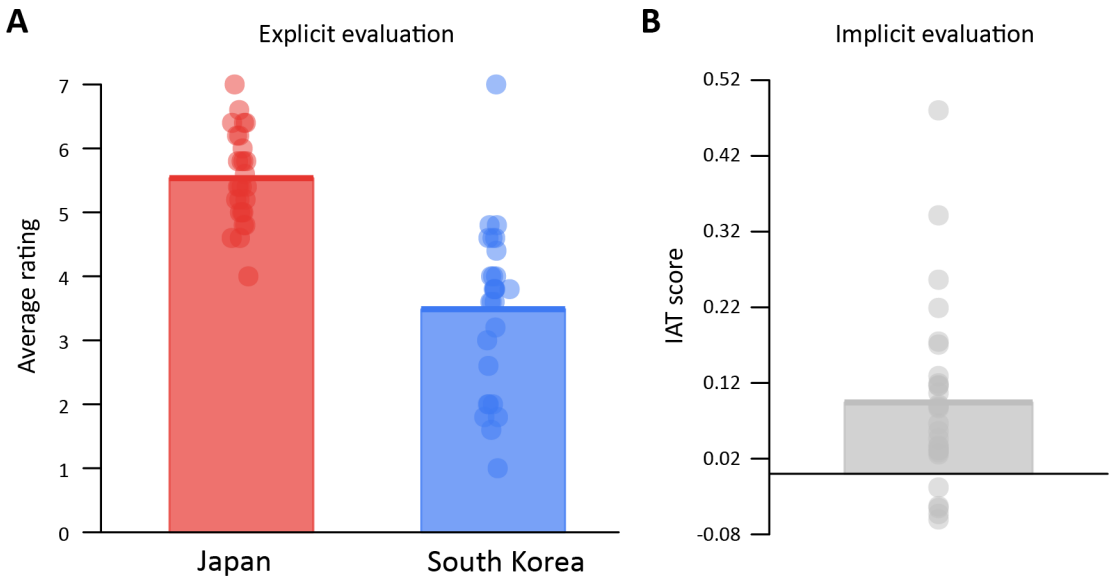


Figure 4.2. (**A**) Bars represent mean explicit evaluations (semantic differentials). Higher numbers indicate more positive evaluation. (**B**) Bar represents mean IAT D-score. Positive scores indicate more positive implicit evaluation of Japan relative to South Korea. Circles denote individual data points.

Effect of Gap on Update

Our multiple regression analyses utilising *Update* as the dependent variable revealed a significant effect of *Gap* (feedback - first estimate) for Japanese ($t(27) = 10.97$ $p < 0.001$) and for South Korean trials ($t(27) = 11.0$ $p < 0.001$), meaning that participants updated their scores *more* from the first to the second rating the larger the gap was between their first rating and the feedback they were presented with. The effect of Favourability was not significant for both Japan and South Korea trials (Table 4.1). However, we observed a significant interaction effect of Gap and Favourability (whether the interaction between the scenario and feedback is overall Favourable or Unfavourable) for Japanese trials ($t(27) = 3.25$, $p = 0.003$) meaning that participants updated their scores significantly more in response to favourable feedback compared to unfavourable feedback. The same interaction effect for the South Korea condition was in the same direction, but didn't reach significance ($t(27) = 1.54$, $p = 0.13$). There was no significant difference in the Gap × Favourability interaction effect between the Japanese and South Korean conditions ($p = 0.30$). Accordingly, although our results showed significantly

more positive implicit and explicit evaluations of Japan compared to South Korea (Figure 4.2, also see Table 4.1), contrary to our prediction, the level of favourability bias is no different between in-group and out-group. Thus, our behavioural results showed that participants tended to update their scores more if the feedback allows them to see other people (regardless of nationality) more favourably. Of final note, it should be stated that no significant difference at group level was observed for any of the Japanese and South Korean predictors (First Estimate $p = 0.23$; Gap $p = 0.68$; Favourability $p = 0.43$; see Table 4.1).

Table 4.1. Behavioural regression model statistics demonstrating beta and *p* values for all predictor variable.

Predictor Variable	Mean Standardised Beta Value	Standard Deviation	*p* value
Japanese			
First Estimate	-7.40	3.84	<0.001**
Gap	7.45	3.60	<0.001**
Favourability	0.67	1.81	0.060
Gap × Favourability	2.06	3.36	0.003**
South Korean			
First Estimate	-8.11	3.72	<0.001**
Gap	7.17	3.45	<0.001**
Favourability	0.28	1.86	0.043*
Gap × Favourability	1.35	4.62	0.134

All values are based on a multiple regression analysis within each participant. P values are based on group level one-sample t-tests. Japanese mean $R^2 = 0.46$, Japanese mean Adjusted $R^2 = 0.42$. South Korean mean $R^2 = 0.44$, South Korean mean Adjusted $R^2 = 0.40$. * $p < 0.05$, ** $p < 0.01$

Correlation of Explicit Attitudes and Favourability Bias Index

Although we didn't observe a significant difference in favourability bias between the in-group and out-group, we observe significant across-subject correlations between explicit evaluations and favourability bias for both Japan ($r = 0.33$, $p = 0.04$) and South Korea ($r = 0.53$, $p = 0.002$), respectively (Figure 4.3). These results are, at least partially, consistent with our prediction and indicate that the strength of favourability bias depends on individuals' attitudes toward a group; the higher the explicit evaluation of Japan or South Korea, the more

participants updated their belief about members of each group when the feedback is in a favourable direction compared to an unfavourable direction.

The Japanese vs. South Korean favourability bias indices were significantly correlated with each other (r = 0.61, p < 0.001), while as stated above, the corresponding explicit evaluations were not significantly correlated with each other (r = 0.16, p = 0.41), indicating that there exists individual differences in viewing other people favourably in general.

Thus, our behavioural results indicate that participants update their ratings more when feedback is in a *favourable* direction as opposed to an *unfavourable* direction, and this effect is seemly consistent across nationalities (Table 4.1). Nonetheless, individual differences in the tendency to update ratings in a favourable direction compared to an unfavourable direction (i.e., favourability bias) were correlated with participants' explicit evaluations for each of the Japan and South Korea conditions (Figure 4.3).

Finally of note, to further examine any bias elicited by participants first estimates, we ran a within-subject correlational analysis to check if participants' first estimates are correlated with Absolute Gap. But, we found no significant correlation for both Japanese (p = 0.32) or South Korean trials (p = 0.38).

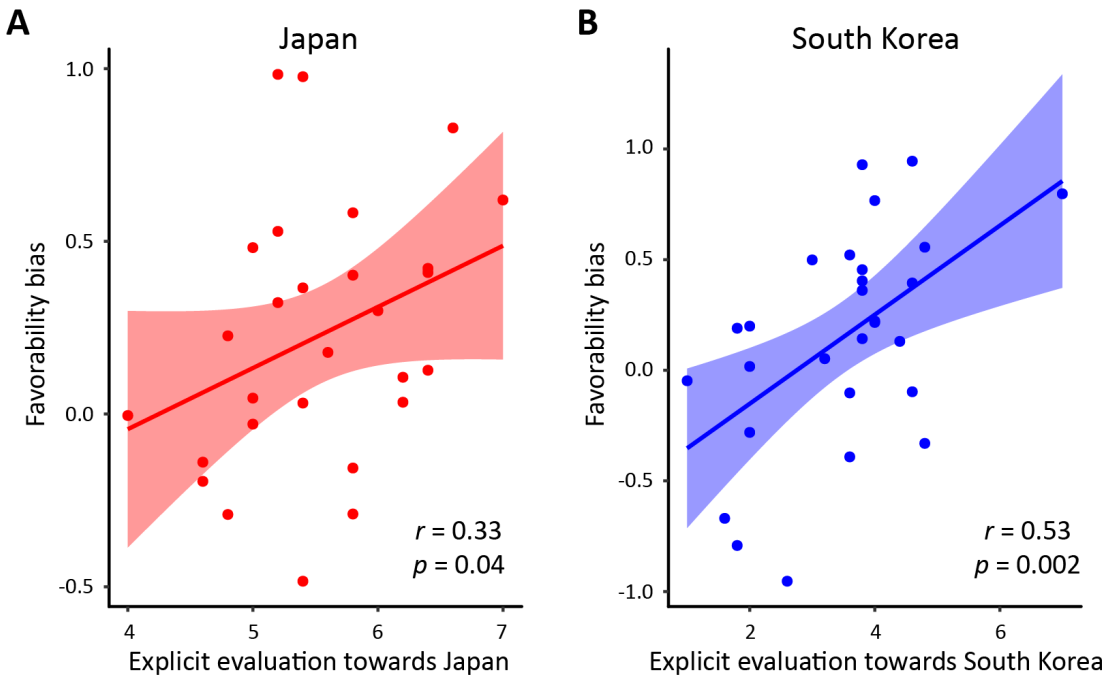


Figure 4.3. Scatter plot demonstrating positive correlation between participants' explicit evaluations of Japan (**A**) and South Korea (**B**), and favourability bias (i.e. the extent participants update their beliefs in favourable trials compared to unfavourable trials). Shaded areas represent 95% confidence intervals.

fMRI Results

Imaging results depicting the effect of Gap

In order to first broadly depict regions related to the conflict between one's initial rating in relation to feedback, we used *Absolute Gap* (absolute value) as a parametric modulator. We investigated the effect of Absolute Gap regardless of condition (i.e., by combining all of the four conditions [Japanese-Favourable, Japanese-Unfavourable, South Korean-Favourable, and South Korean-Unfavourable]). Here, we found that pMFC (specifically the dmPFC and left supplementary motor area; SMA), lateral superior temporal gyrus (STG), and posterior cingulate cortex (PCC) activity is positively correlated with Absolute Gap (see Table 4.2 & Figure 4.4A, B, & C). These regions are largely consistent with areas previously implicated in social conflict (the difference between one's and others' opinions) in a social conformity paradigm (Izuma & Adolphs, 2013; Klucharev et al., 2009; Wu et al., 2016). For full information of the overlap between the current studies activation map and that of Izuma and

Adolph (2013), see Supplementary Results (Appendix 2). In our main ROI of the dmPFC (x = -8, y = 24, z = 66), the effect of Gap was significantly positive in all conditions excluding Japanese Favourable, which was marginally insignificant (Japanese Favourable $p = 0.08$, all remaining $ps < 0.001$; Figure 4.4C).

Furthermore, examination of brain regions *negatively* correlated with Absolute Gap revealed significant activation within the ventral striatum (specifically nucleus accumbens, see both Table 4.2 for full list of regions activated and Figure 4.5A & B for associated contrast image), also consistent with previous studies. For results of regions correlated with Absolute Gap for each condition separately (Japanese-Favourable, Japanese-Unfavourable, South Korean-Favourable, South Korean-Unfavourable), see Supplementary Table 4.S4.

Table 4.2. Brain regions correlated with Absolute Gap

Location	BA	MNI coordinate			Z	Cluster size
		x	y	z		
Areas *positively* correlated with Absolute Gap						
dmPFC	8	-8	24	66	5.16	1996
left supplementary motor area (SMA)	8	-6	22	58	5.12	
left superior frontal gyrus (SFG)	9	-12	46	46	4.87	
Right superior temporal gyrus (STG)	20	44	16	-36	4.84	327
Left superior temporal gyrus (STG)	30	-42	20	-30	5.07	1569
left pars orbitalis gyrus	47	-44	32	-6	4.89	
left insula	47	-40	22	-8	4.87	
Posterior cingulate cortex (PCC)	23	-6	-50	28	4.80	1076
Areas *negatively* correlated with Absolute Gap						
Right postcentral gyrus	40	56	-40	50	6.18	2321
right supramarginal gyrus	40	46	-36	40	5.60	
right angular gyrus	40	40	-48	54	5.07	
Left postcentral gyrus	40	-48	-36	44	5.82	2431
left angular gyrus	40	-54	-40	54	5.79	
Right middle frontal gyrus (MFG)	8	26	16	56	5.98	557
Left middle frontal gyrus (MFG)	46	-38	34	26	5.55	931
Right ventral striatum	25	12	10	-10	5.42	1051
right pars opercularis gyrus	44	52	12	24	4.98	

BA, Brodmann area. Statistics are based on a set threshold of height $p < 0.001$ (uncorrected), and cluster $p < 0.05$ (FWE). Areas in grey italics represent significant peak (FWE) sub-clusters/different regions within larger clusters.

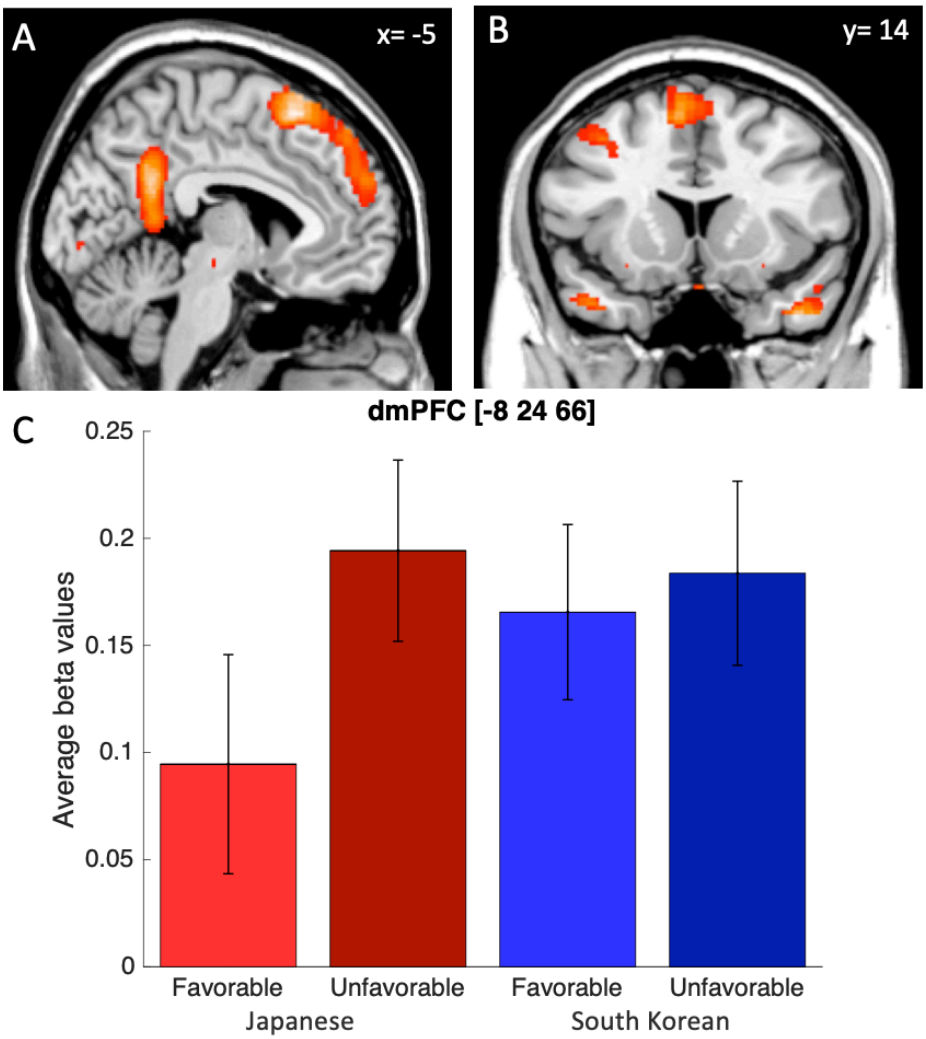


Figure 4.4. (**A)** Sagittal slice (x = -5) demonstrating brain regions positively correlated with Absolute Gap. (**B**) Coronal slice (y = 14) demonstrating brain regions positively correlated with Absolute Gap**. (C**) Bars represent average beta values across all conditions within key significant cluster in the dmPFC, error bars denote SEM. All betas were extracted via a 4mm sphere from the peak activation identified by the contrast image depicting all trials modulated by Absolute Gap.

Interestingly, exploration of the contrast image depicting activation for Unfavourable trials modulated by Absolute Gap compared to Favourable trials modulated by Absolute Gap (Unfavourable > Favourable) also revealed that a different cluster within the dmPFC (x = 6, y = 38, z = 48, k = 238), left inferior frontal gyrus (IFG, x = -48, y = 18, z = 22, k = 1137) and right middle frontal gyrus (MFG, x = 40, = 8, z = 58, k = 607) was more sensitive to Absolute Gap in an *Unfavourable* direction compared to a Favourable direction (see Figure 4.6A & B). As shown in Figure 4.6C, the dmPFC tracked Absolute Gap in an Unfavourable direction,

while it was insensitive to Absolute Gap in a Favourable direction. In contrast, no clusters survived the threshold in place when examining brain regions correlated with Absolute Gap for Favourable trials compared to Unfavourable trials (for full list of results, see Supplementary Results, Appendix 2, Supplementary Table 4.S5).

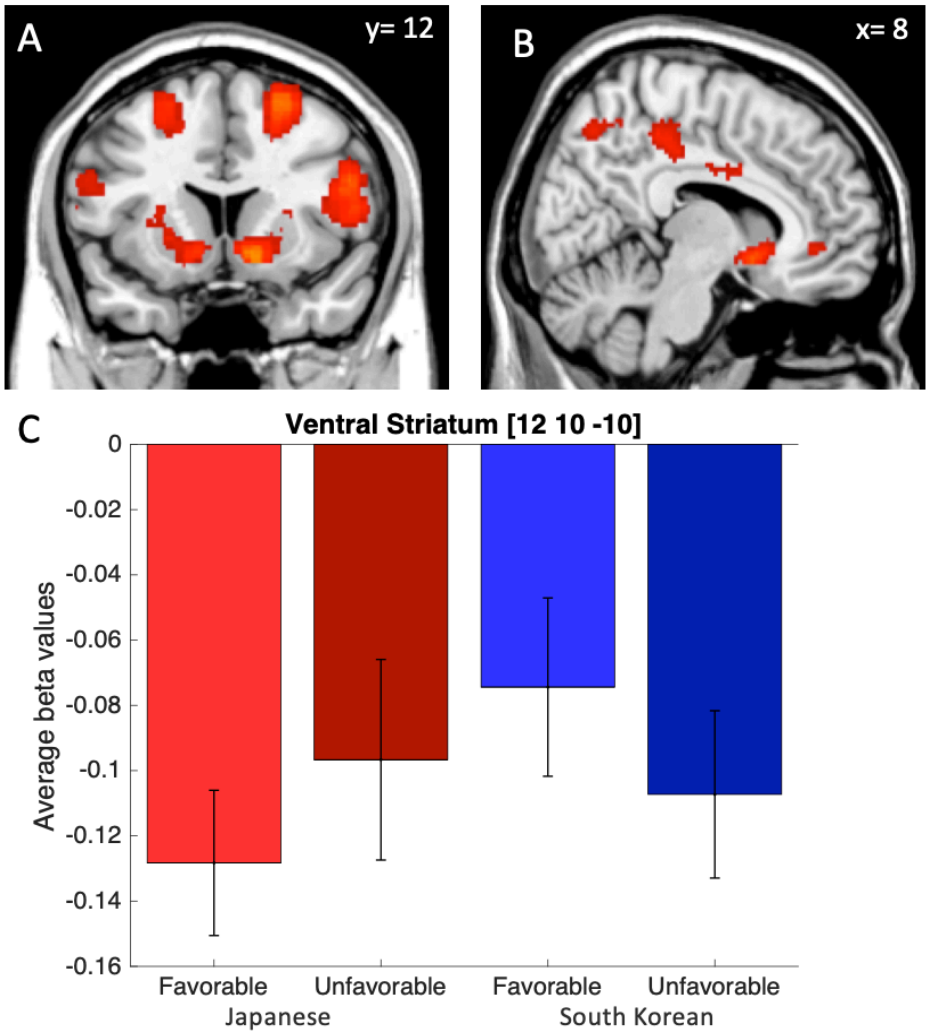


Figure 4.5. (**A**) Coronal slice (y = 12) demonstrating brain regions negatively correlated with Absolute Gap. **(B)** Sagittal slice (x = 8) demonstrating brain regions negatively correlated with Absolute Gap. (**C**) Bars represent average beta values across all conditions within key significant cluster in the ventral striatum. All betas were extracted via a 4mm sphere from the peak activation identified by the contrast image depicting all trials modulated by Absolute Gap, and error bars denote SEM.

We additionally explored several brain-behaviour correlations. Although our behavioural results revealed robust individual differences in favourability bias, there was no significant correlation between the behavioural favourability bias and neural favourability bias (i.e.,

Unfavourable-Absolute Gap vs. Favourable-Absolute Gap) in the dmPFC (or any additional ROIs reported in Table 4.2) for both the Japan (r = 0.21, p = 0.29) and South Korea (r = -0.00, p = 0.98) conditions.

Thus, while our behavioural data showed that participants' updated their estimates more when the feedback was in a *favourable* direction, our fMRI data actually indicated that the cluster within the dmPFC (x = 6, y = 38, z = 48; Figure 4.6A) was more sensitive to the discrepancy between one's initial estimate and the feedback when the feedback was in an *unfavourable* direction.

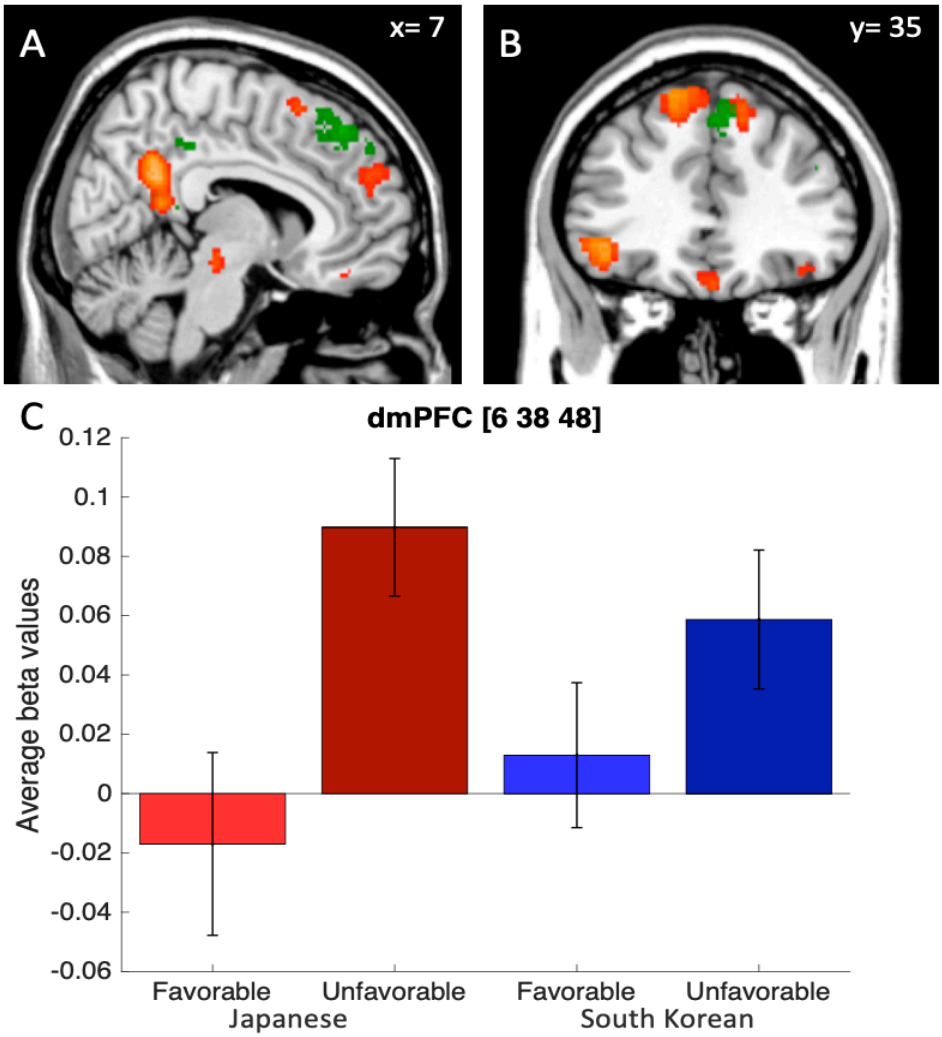


Figure 4.6. (**A**) Sagittal slice (x = 7) demonstrating brain regions correlated with Absolute Gap (all of the four conditions combined; shown in orange), as well as brain activity for unfavourable compared to favourable trials modulated by Absolute Gap (shown in green). This contrast partially replicates Figure 4.4A (activation shown in orange) from a slightly different slice perspective in order to demonstrate the independent nature of the dmPFC sensitivity specifically for unfavourable trials (green) compared to across all trials (orange). (**B**) Coronal slice (y = 35) demonstrating brain regions correlated with Absolute Gap (all of the four conditions combined; shown in orange), as well as brain regions significantly more strongly correlated with Absolute Gap in unfavourable trials compared to favourable trials (shown in green). (**C**) Bars represent average beta values across all conditions within key significant cluster in the dmPFC (x = 6 y = 38 z = 48). All betas were extracted via a 4mm sphere from the peak activation identified by the contrast image depicting unfavourable compared to favourable trials modulated by Absolute Gap. All error bars denote SEM.

Imaging Results depicting the effect of Update

In order to further assess whether any brain regions are related to the actual *change* of participant's ratings (Update), the same parametric modulation analysis was conducted using

Update (controlled for RTM) as the parametric modulator, instead of Absolute Gap. No significant clusters survived the threshold, and although alluded to in some previous research regarding the pMFC and attitude change, no significant activation in these regions were observed via the same contrast image combining all conditions modulated by Update.

DISCUSSION

The aim of the study was to test two competing hypotheses regarding pMFC activity, those being; if the pMFC encodes the conflict between reality and a desirable outcome, or if the pMFC plays a role in belief updating. This was assessed by employing a cognitive bias to specifically disentangle the level of conflict from the level of belief updating, whilst assessing pMFC sensitivity respectively. Accordingly, our behavioural data indicates participants' are more likely to update their beliefs in the direction of *favourable* new information (especially in the Japan condition), whilst our fMRI data indicates that the dmPFC is more sensitive to *unfavourable* new information (Figure 4.6A), and this effect was consistent across Japanese and South Korean conditions. In contrast, no brain region was significantly related to behavioural update. Thus, the findings support the conflict hypothesis rather than the update hypothesis, indicative that sensitivity in the pMFC (at least within the dmPFC; Figure 4.6A) is related to the conflict between ideal scenarios versus reality.

Activation of the dmPFC in Izuma and Adolph (2013) tracked the discrepancy between one's own preference and its social ideal as defined by balance theory (Heider, 1946). In the current study we see a matching activation map to that of Izuma and Adolph (2013) across all combined conditions modulated by Absolute Gap (basically the degree of conflict in each trial, hereby referred to as such for the purpose of the discussion) (Figure 4.4). However, the same neural activation in regards to solely the updating of beliefs based on new information was not observed. Henceforth, it would seem likely that brain activity demonstrated in the current

experiment is liable representative of the conflict of information presented, rather than any associated updating of beliefs. Nonetheless, it should be specified that the analysis is based on the onset of feedback presentation, not when participants give their second estimates, where any additional neural mechanisms (potentially the dmPFC) related to the update of belief may be more apparent. Although we focused on brain activations during the feedback processing in the first rating task just like a majority of previous social conformity studies, this might explain why under the current paradigm, no significant neural activity regarding the updating of beliefs was seen. It is interesting and important to see in future research whether the dmPFC, or other brain regions, tracks the degree of behavioural adjustments (update) similar to the ones implemented in the current study during the second rating task.

A key result from this study was that the dmPFC, left IFG, and right MFC were more sensitive to the degree of conflict in unfavourable compared to favourable trials. This tallies with Holroyd and Cole (2002), who highlight the pMFC's involvement with the focus on consequence predication in terms of action monitoring, specifically, when the outcome of a given task is *worse than expected*. An effect also relevant to this paradigm is the "False Consensus Effect" (Ross et al., 1977), the notion that people tend to believe more people share their attitudes/world view than actually do. Interestingly, Welborn and Lieberman (2018) found when examining the neural effects of consensus bias, pMFC (specifically the medial prefrontal cortex and ventral medial prefrontal cortex: mPFC, vmPFC) activity was positively associated with observed consensus bias only when information given to participants as feedback (similar to this study) was of a challenging/disconfirmatory nature, as opposed to confirming previous beliefs. Thus, our work appears to replicate a specific sensitivity of goal-driven conflict within the pMFC, also fitting nicely with a recent review regarding the motivational characteristics of cognitive consistency, that being we strive more for specifically *favoured* outcomes rather than consistent ones alone (Kruglanski et al., 2018).

Although the present study demonstrated that these regions were more sensitive to unfavourable information, it was favourable information that was more successfully updated in the second rating task. The contrast between our fMRI and behavioural data on the surface resembles the general effect of cognitive dissonance (discomfort evoked by the discrepancy between attitudes, beliefs, and behaviour) (Festinger, 1962), a form of conflict in its simplest form. That being, participants seemingly exhibit more negative emotion from the unfavourable feedback (indicated by increased sensitivity in the aforementioned ROIs), yet do not update it as efficiently. This allies with previous research which also posits the pMFC (Harmon-Jones et al., 2008) as being a central neural correlate of cognitive dissonance, particularly in the dmPFC (Izuma et al., 2010) and dACC (Izuma et al., 2010; Van Veen et al., 2009; Izuma & Murayama, *in press*). However, it should be said that in more typical examples of cognitive dissonance, participants often resolve this by amending behaviour and/or attitudes accordingly, whereas in the current study participants seem to resolve this conflict by not (or to a lesser extent) updating their belief according to unfavourable information (further discussion on the lack of memory update is extended in the next paragraph). One important distinction to first make here is that participants' also have an additional conflict of being "correct", since there is a factually correct answer in this experimental paradigm, whereas classic cognitive dissonance studies tend to revolve around preference (which participants can freely change). This avoids any extra level of divergence the current participants' may have underwent (resolving dissonance vs. being correct), which could possibly have added to the lack of update observed in the current experiment.

Relatedly, and in somewhat contrast to the current study, Hughes et al., (2017) found participants were more likely to update their impressions regarding negative information during an impression formation task about out-group members, but not in-group members. This was associated with less engagement in the dACC, temporoparietal junction, insula, and

precuneus when processing negative information about the in-group, but importantly not the out-group. The asymmetry of participants impression update and neural response between in versus out-group members suggests that these neural structures are important for updating one's impression, especially when new information fits with individual's pre-existing notions (e.g., in-group positive behaviour and out-group negative behaviour). Though this study is similar in many ways to the current experiment, there are several key differences. First relates to the point above regarding the re-assessment of subjective (opinion) versus objective (facts) information, which is an important distinction between Hughes et al., and the current study. Second, it should also be noted that though we do measure subjective impressions (explicit attitudes) of the out-group (and in-group) as they do in Hughes et al., (2007), because this was only measured at a single timepoint in the current experiment, it isn't possible to compare any possible update/change of this after participants received feedback. Finally, it's also relevant to highlight that the participants who produced lower explicit attitudes towards the out-group did tend to update more unfavourable information, allying with Hughes et al., (2017) findings.

In order to continue to elucidate the role of the dmPFC, it is increasingly important to assess the effect of memory. In an apparent contrast to our results, previous research would suggest that more conflicting or shocking information is more likely to be remembered (Berntsen, 2002; Kensinger, 2007). This might suggest that unfavourable information was not updated due to participants' active inhibition of the effect of unfavourable information on update during the second estimation task. Alternatively (but not necessarily mutually exclusive), what may be apparent is inefficient encoding of the feedback during the first estimation task. Our data demonstrates that activity in the left IFG, and the dmPFC was more sensitive to Gap in unfavourable trials compared to favourable trials (Figure 4.6), and these two regions have been implicated in response inhibition (Floden & Stuss, 2006; Verfaellie & Heilman, 1987). Historically, increased activation in the right (as opposed to the left) IFG has

been associated with increased inhibitory control of responses (e.g. De Zubicaray et al., 2000; Garavan et al., 1999; Konishi et al., 1999), but there is some suggestion that the left IFG also plays a central role in response inhibition. Specifically, Swick et al., (2008) found patients with left IFG legions had higher error rates than controls in both conditions (easy vs. hard) of a Go/NoGo task, being further impaired in the hard condition when more inhibitory control was required. Future research should examine more extensively neural activities during the second rating task and the relationship regarding the valence of social information and subsequent memory processes (e.g., whether unfavourable feedback is better remembered) to tease apart the two possibilities (increased inhibition vs. decreased encoding).

Further ROIs we found from the fMRI data include areas of the striatum (nucleus accumbens specifically) which were negatively correlated with the degree of conflict in each trial (Figure 4.5). This supplements previous research that also demonstrates when participants' opinions differ from that of others, whilst the pMFC is activated, the striatum is deactivated (Campbell-Meiklejohn et al., 2010; Izuma & Adolphs, 2013; Klucharev et al., 2009). Welborn and Lieberman (2018) infer their similar finding in terms of the gratifying value of information. This seems a tenable explanation, with additional links made toward reinforcement learning surrounding conformity by Klucharev et al., (2009). Alternativly, it seems an important distinction that our dmPFC (Figure 4.4) and ventral striatum clusters encode Absolute Gap across all trials (positively: dmPFC, or negatively: ventral striatum) in a relatively objective manner (i.e., unaffected by favourability of information), suggesting these regions are related to general learning mechanisms. On the other hand, the dmPFC cluster that encodes Absolute Gap specifically for Unfavourable compared to Favourable trials (Figure 4.6) seems to be influenced by a top down emotional process so that in addition to the objective difference (Absolute Gap), the activity is modulated by what participants *hope* the reality to be. Thus, our ventral striatum activation may represent the processing of information more objectively

(rather than subjectively being influenced by the valuation of information). This relates nicely to a recent fMRI study by Pine et al., (2018), which specifically highlights the ventral striatum's involvement in the learning of factual knowledge.

Our results also demonstrate increased sensitivity for the degree of conflict within the PCC and lateral STG. The PCC has been implicated in tracking the cognitive imbalance between own preferences versus others, as well as being correlated with subsequent preference changes in Izuma and Adolph (2013). Furthermore, work by Falk et al., (2014) show the PCC is more sensitive to social exclusion in participants who also subsequently change their actions to suit peers (in this case, increase the level of risk in their driving more around peers as opposed to alone). Although our data doesn't demonstrate an association with the behavioural update, it seems consistent that this region plays a role in the recognition of social conflict. Not only has this been established in terms of social conflict (see also Seehausen et al., 2014), neuroimaging studies have also shown the PCC to be sensitive in monitoring non-social prediction errors and conflict in general (Christoffels, Formisano, & Schiller, 2007; Kadosh, Kadosh, Henik, & Linden, 2008). The STG has some similar implications in the monitoring of social conflict (Christoffels et al., 2007). For example, Premkumar et al., (2012) report the right STG to be more active during the viewing of social rejection as opposed to neutral scenes, and Seehausen et al., (2014) found the STG to be more active in an empathy-experiment where participnats felt misunderstood (in comparison to understood)- both implicating a potential role in the discrimination of desirable versus undesirable outcomes.

Behaviourally, participants demonstrated a favourability bias in general. We display a correlation between positive evaluations to Japan *or* South Korea and the extent participants update their beliefs based on more favourable information. More broadly put, participants increasingly revise their belief based on new information to see people more positive for previously more liked social groups, supplementing the previously discussed work of Izuma

and Adolph (2013). As participants overall possessed positive explicit evaluations of Japan, the data coincides with our behavioural hypothesis that more beliefs are updated regarding favourable information. However, although our participants explicit and implicit evaluations were on average significantly less positive for South Korea, participants did still elicit a favourability bias at the group level for South Korea also, updating their beliefs more so for favourable trials here too.

Our initial behavioural hypothesis stated that any favourability effect would be less pronounced for South Korea owing to less positive attitudes in general. This outcome was forecast to arise due to the effect of confirmation bias, seeing participants update information that more aligns with their previous attitudes (more positive towards Japan versus less positive towards South Korea). An initial consideration here, then, is that the results are more consistent with the "good-news-bad-news-effect" (Eil & Rao, 2011). This is the concept that information and its corresponding valence are not updated and processed in an equal, linear manner. Positive information (good news) tends to revise according to previous experience and is more efficiently updated, whereas the updating of negative information (bad news) is not, being more noisy and less likely to be updated into current beliefs. Broadly applied to the current findings, this would suggest that updating favourable compared to unfavourable information takes place in a more efficient and uniform manner, regardless of any pre-existing views and thus the social group applied to. This has been supported by work on optimism bias (Sharot et al., 2011), demonstrating participants' are more likely to update their belief based on more positive information about the future compared to negative information. This positivity bias is theorised to arise as a protection for general mental well-being (Garrett et al., 2018; Sharot et al., 2011).

It should also be noticed that the explicit evaluations towards South Korea displayed large across-participant variability, with many participants having close-to-neutral attitudes (meaning they didn't feel particularly positive *or* negative towards South Korea). But to

reiterate, the participants who *did* have extremely negative explicit evaluation's towards South Korea did tend to update their beliefs more in response to *unfavourable* feedback. Speculatively, since we only measured explicit attitudes at a single time point, these results might suggest that more moderate attitudes are increasingly amendable upon receiving information, more easily disconfirming any pre-existing *weaker* stereotypes. This, in comparison to more extreme attitudes in which the information may be updated more asymmetrically (as presented by Sunstein et al., 2016), further facilitating attitude polarisation, additionally coincides with research that demonstrates increased dogmatic-intolerance in relation to attitude extremity (van Prooijen & Krouwel, 2017).

Future research may wish to select a more exclusively hostile and defined in/out-group paradigm in order to further extract any additional effects of attitude extremity, and the associated neural correlates/behavioural update. For example, it may be interesting to examine a potential ceiling (or cross-over) effect of the good-news-bad-news model in terms of extreme attitudes- at what point is bad news about a disliked out-group no longer perceived as "bad", but instead information that only affirms ones previous distain? What's more, if the pMFC is sensitive to social conflict as we showed, this should in theory then be less robust for negative information regarding disliked out-groups for people with extremely negative attitudes due to lesser conflict between ones social outlook versus reality. Finally, although we found similar neural correlates of Absolute Gap (Figures 4.4 & 4.5) between the present study with Japanese participants and our previous study with American participants (Izuma & Adolphs, 2013), it is important to systematically and directly test cultural differences in social information processing in future research, as previous studies indicate cultural differences in social conformity (Bond & Smith, 1996; Korn et al., 2014) and cognitive dissonance (Kitayama, Snibbe, Markus, & Suzuki, 2004 but see also Chen & Risen, 2010; Izuma & Murayama, 2013).

CONCLUSION

In sum, the current experiment demonstrated two key points, i) activity in the dmPFC was representative of socially conflicting information, specifically the conflict between ideal outcomes versus less ideal realities, and not the corresponding belief update based on new information. ii) participants updated their beliefs based on more favourable information, of which related to more positive evaluations of the social group in question. Future research should aim to further disentangle the role of the dmPFC in social conflict processing, attempting to apply experimental paradigms to specifically isolate potentially independent neural correlates related to the actual update of participants beliefs based on new information received. What can be taken from the current study overall is an increased understanding of the role played by the dmPFC in social information processing, of which ultimately helps us to understand how decisions about social interactions are made, providing a more solid foundation for social attitude amendment and interventions.

Chapter 5

Specialised Social Mechanisms

Chapters two, three, and four of the current thesis are centred around the various neural mechanisms underlying the processing of inconsistent and conflicting social information. The final empirical chapter will relate to the premise of a specialised network for the processing of social information.

A network of brain regions specifically adapt to process social information and underlie social cognition, coined the 'social brain', originally comes from Brothers (1990). It predominantly encompasses the amygdala, orbital frontal cortex and temporal cortex as its key components. This network is heavily involved in governing social cognition (Apperly, 2010; Saxe, 2010), emotion (Reeck, Ames, & Ochsner, 2016), and behaviour (Montague & Lohrenz, 2007; for a recent review see Ugazio & Ruff, 2017). Several evolutionary accounts offer conceptual support to this notion, for example Dunbar's (1998) Social Brain Hypothesis. This is the idea that as social group size increased in our evolutionary history, interactions became more complex and required more sophisticated levels of social networking in order to thrive. This is demonstrated across primates by observing the positive relationship between social group size and brain size (Dunbar & Shultz, 2007a, 2007b), and also a potential limit to group size dependant on the sophistication of information processing across organisms (Dávid-Barrett & Dunbar, 2013).

Research has shown distinct neural activation for social versus non-social mechanisms, for example Martin and Weisberg (2003) found when participants viewed geometric animations that purposefully conveyed social interaction (relative to conveying mechanical interaction) it elicited activation in regions of the posterior temporal cortex, previously associated with identifying human faces (Haxby, Hoffman, & Gobbini, 2000; but see also Duchaine & Yovel, 2015). Yet, it remains not yet clarified whether the brain has a specific

neural circuit for exclusively social interactions, or if this network is a product of information processing that can also be relevant to non-social information. In terms of decision making, Ruff and Fehr (2014) proposed two schematic processes for dealing with social versus non-social stimuli, the first being the "extended common currency schema" which argues similar neural processes assign motivational relevance to social/non-social information, suggesting a similar network of brain function that encode social versus non-social information. Secondly, the "social valuation specific schema" assumes a devoted neural network which specifically encode values accompanying interactions and decisions that involve others.

In order to investigate the parallels between social and non-social information processing, it seems important to isolate a factor that can be viewed as both social and non-social. One example of this is the emotion of reward. This can be in terms of material gain (i.e. money, prizes) or social (i.e. compliments, increased reputation).

Chapter 6

A common neural code for social and monetary rewards in the human striatum

Stephanie J. Wake[1] and Keise Izuma[1,2]

[1] Department of Psychology, University of York, Heslington, York, YO10 5DD, UK.

[2] Department of Psychology, University of Southampton, University Road, Southampton, SO17 1BJ, UK

ABSTRACT

Although managing social information and decision making on the basis of reward is critical for survival, it remains uncertain whether differing reward type is processed in a uniform manner. Previously, we demonstrated that monetary reward and the social reward of good reputation activated the same striatal regions including the caudate nucleus and putamen. However, it remains unclear whether overlapping activations reflect activities of the same neuronal population or two overlapping but functionally independent neuronal populations. Here, we re-analysed the original data and addressed this question using multivariate pattern analysis (MVPA) and found evidence that in the left caudate nucleus and bilateral nucleus accumbens, social versus monetary reward were represented similarly. The findings suggest that social and monetary rewards are processed by the same population of neurons within these regions of the striatum. Additional findings also demonstrated similar neural patterns when participants experience high social reward compared to viewing others receiving low social reward (potentially inducing schadenfreude). This is possibly an early indication that the same population of neurons may be responsible for processing two different types of social reward (good reputation and schadenfreude). These findings provide a supplementary perspective to previous research, helping to further elucidate the mechanisms behind social versus non-social reward processing.

Key Words: social reward, monetary reward, schadenfreude, fMRI, striatum, MVPA

INTRODUCTION

Consider this; i) people think you are wonderful and regard you as a great person, ii) You win a £100 prize in a raffle. Both feel good, but it remains uncertain whether social reward and non-social tangible reward share the same neural mechanisms. Making important decisions that dictate survival based on both social and non-social information is a part of everyday life, yet we know relatively little about the comparative reward types that we seek on a daily basis.

An abundance of neuroscience studies has found various social and non-social rewards activate the striatum (Fehr & Camerer, 2007; Izuma, 2015). It is well established in non-human neurophysiological studies that striatal neurons respond to reward (Schultz, Tremblay, & Hollerman, 2000), and this basic finding has been later replicated by human neuroimaging studies (Delgado, 2007). More recently, social neuroscience and neuroeconomics studies demonstrated that the striatum is activated by a variety of socially rewarding stimuli or behaviour, such as mutually cooperating with other individuals (Rilling et al., 2002; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004), punishing unfair behaviour (De Quervain et al., 2004; Singer et al., 2006), giving charitable donations (Harbaugh, Mayr, & Burghart, 2007; Moll et al., 2006) and receiving a good reputation from others (Izuma et al., 2008; Korn, Prehn, Park, Walter, & Heekeren, 2012).

An important question, which remains unanswered in the field, is whether social and non-social rewards share a common neural mechanism. Importantly, activation overlaps between social and non-social rewards reported previously (Fehr & Camerer, 2007; Izuma, 2015) cannot be taken as strong evidence for a shared neural mechanism. It may indeed reflect the same population of neurons responding to both types of rewards (i.e., a shared neural mechanism) or it could in fact signify largely distinct populations of neurons specialised for each reward, which are located in close proximity within the same brain region (e.g., striatum). Ruff and Fehr (2014) proposed two schematic processes for dealing with social versus non-

social stimuli. The first being the "extended common currency schema", which argues identical neural processes assign motivational relevance to social/non-social information, predicting similar populations of neurons that encode reward values of both social and non-social stimuli. Secondly, the "social valuation specific schema" assumes an evolved and dedicated neural circuitry which specifically encode reward values associated with interactions and decisions that involve others. This predicts that there are distinct populations of neurons that process social and non-social rewards.

In the present study, we aim to provide an insight into this question by applying multivariate pattern analysis (MVPA) (Norman, Polyn, Detre, & Haxby, 2006) to the data reported previously (Izuma et al., 2008). In the original study (Izuma et al., 2008), the same participants were asked to perform tasks involving non-social reward (money) and social reward (good reputation from others), and found that the striatum (see Figure 6.1; especially the left putamen and left caudate nucleus) were significantly activated for both monetary and social rewards. Using MVPA, the present study further investigates whether the pattern of activity across multiple voxels within the striatum is similar between social and monetary rewards (i.e., that social and monetary rewards share common neural networks).

How to interpret activation overlaps has been a recurring question in cognitive neuroscience, and MVPA is a useful tool that allows us to infer activities of underlying neuronal populations from fMRI signals, helping us interpret the overlaps (Kaplan, Man, & Greening, 2015; Peelen & Downing, 2007). For example, Woo et al., (2014) found physical pain and social pain, previously known to activate the same regions within the dorsal anterior cingulate cortex (dACC) and insula (Kross, Berman, Mischel, Smith, & Wager, 2011), actually showed distinct activation patterns under MVPA, providing important evidence against a popular notion that physical and social pain share the same neural representation (Eisenberger, 2012). Similarly, using MVPA, Krishnan et al., (2016) found that felt and seen pain, also

known to activate the same dACC region (Singer et al., 2004), in fact demonstrate distinct activation patterns. Thus, as these overlaps that were once thought to indicate a similar neural mechanism under conventional univariate analysis are actually found to be discriminate under MVPA, it seems fundamental that fMRI research utilise this technique to further assess whether underlying neuronal populations are similar.

MATERIALS and METHODS

Participants

Data from 19 participants (9 male; mean age = 21.6 ± 1.5 years) were included in the reanalysis using the existing dataset (Izuma et al., 2008). All participants gave written informed consent for participation, and the study was approved by the Ethical Committee of the National Institute for Physiological Sciences, Japan.

Procedure

Full details for procedures used in the study have been published previously (Izuma et al., 2008). Briefly, each participant completed two different fMRI experiments (involving monetary and social rewards, respectively) on two separate days.

In the first monetary reward experiment, participants took part in a simple gambling task. In each trial, they were asked to choose one of three cards and were given 0, 30, or 60 yen depending upon the card chosen. However, the amount that they could earn in each block of eight trials was predetermined; thus, the monetary reward each participant received during each block was systematically manipulated. There were three reward levels (i.e., conditions); 1) High, 2) Low, and 3) No reward (control). After the monetary reward experiment, participants were asked to respond to several personality questionnaires and to introduce themselves in front of a video camera. Participants were specifically told that others would evaluate them

based on their responses to these questionnaires and the video-taped self-introduction, and that they would be shown the results in the next fMRI experiment.

In the second social reward experiment, the same 19 participants were presented with a picture of themselves and a word or phrase indicating the impression of them formed by others. In reality, the items presented were predetermined, such that all participants had the same social reward experience. By systematically grouping six items (into one block) based on desirability ratings provided by another group of participants (n = 33), the level of social reward experienced by participants in each block was also manipulated. To exclude the possibility that seeing a positive word per se might be rewarding, as was suggested by a previous study (Hamann & Mao, 2002), the impressions of other people were also presented. Thus, there were six conditions in the second experiment (a 2 [Target; Self or Others] × 3 [Reward level; High, Low or No reward] within-subject design).

Data Analysis

fMRI data was re-analysed using SPM8 as implemented in Matlab 8.1. Head motion was corrected using the realignment program, and the volumes were normalised to the Montreal Neurological Institute (MNI) space using the EPI template (resampled voxel size 2 × 2 × 2mm). Spatial smoothing was not applied in order to preserve fine grained activation patterns for multivariate analyses.

Correlation-based MVPA: As done in the original correlation-based MVPA study (Haxby et al., 2001), the data for each participant was split into odd versus even runs. This is mainly intended to check the within-condition correlation as well as to get an insight into whether the same population of neurons process social and monetary rewards. For example, if a striatal region processes information related to monetary or social reward, the same condition (e.g., High Monetary Reward condition) should evoke similar activation patterns across

different runs (i.e., significant within-condition correlation). Similarly, if the same population of neurons encode social and monetary rewards, the two conditions should evoke similar activations patterns (i.e., significant between-condition correlation). It should be noted that using the average absolute values of the difference in each realignment parameter between one scan and its successive scan as a motion index (e.g., Yoo, Choi, Juh, Pae, & Lee, 2005 Neuroscience Research), we confirmed that there was no significant difference in head motion (in each of the six motion parameters) between odd vs. even runs in both monetary and social experiments (all *p*s > 0.103).

Since each of the monetary and social reward experiments had four fMRI runs, we conducted the same first level analysis using a general linear model as our original study (Izuma et al., 2008), and contrast images were generated separately for odd and even runs, yielding a total of 18 contrast images for each participant; 6 contrast images from the monetary reward experiment (2 [fMRI Run; odd or even] × 3 [Reward level; High, Low or No reward], and 12 contrast images from the social reward experiment 2 [fMRI Run; odd or even] × 2 [Target; Self or Others] × 3 [Reward level; High, Low or No reward]). These 18 contrast images were used in the correlation-based MVPA.

Using the data from each of the four regions of interest (ROIs; see below), correlation-based MVPA computes a voxel-by-voxel correlation between one condition in odd runs and the same (within-condition correlation) or different (between-condition correlation) conditions in even runs within each participant. The resulting correlation values are fisher-z transformed and submitted to group level analyses (i.e., one-sample t test [one-tailed]).

Classifier-based MVPA: To check the robustness of our results (especially in the left caudate nucleus), we also ran classifier-based MVPA (a linear support vector machine), which was performed by using custom-made Matlab scripts in combination with LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). For this analysis, contrast images for each of the

four fMRI runs were created separately, and classification performances were evaluated by a leave-one-run-out cross-validation procedure. We first trained and tested a classifier that discriminates the High Monetary Reward condition from the No Monetary Reward condition (i.e., monetary reward classifier). Similarly, we next trained and tested a classifier that discriminates the High Social Reward-Self condition from the No Social Reward-Self condition (i.e., social reward classifier). Finally, we tested whether the monetary reward classifier can discriminate the High Social Reward-Self condition from the No Social Reward-Self condition, and similarly whether the social reward classifier can discriminate the High Monetary Reward condition from the No Monetary Reward condition, an approach known as Multivariate Cross-Classification (MVCC; Kaplan et al., 2015).

Regions of Interest (ROI): Striatal areas commonly activated by both monetary and social rewards, which were reported in the original study (Izuma et al., 2008), included the caudate nucleus and putamen bilaterally. Thus, in order to limit each MVPA to the same anatomical region, we applied anatomical masks (the WFU PickAtlas toolbox for SPM; Maldjian, Laurienti, Kraft, & Burdette, 2003) to the original activation map and created four ROIs (see Figure 1); 1) right caudate nucleus (125 voxels), 2) left caudate nucleus (87 voxels), 3) right putamen (110 voxels), and 4) left putamen (99 voxels).

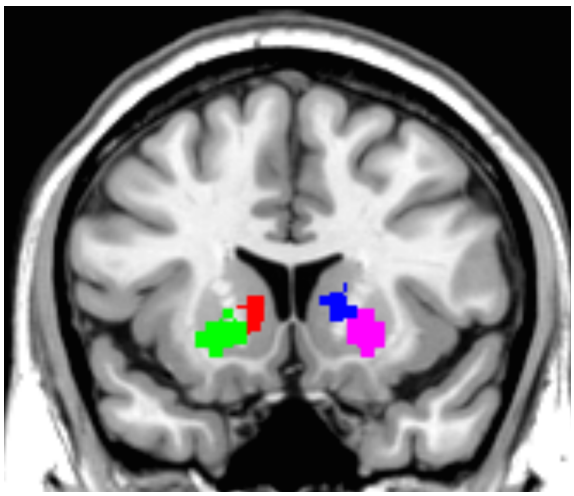


Figure 6.1. Axial slice (y = 14) showing the four ROIs used in the MVPA. These four regions were commonly activated during social vs. monetary rewards in the original study (Izuma et al., 2008).

Exploratory Searchlight Analysis: In addition to the ROI based MVPA mentioned above, we conducted a searchlight MVPA to explore whether any other regions within the striatum represent social and monetary rewards in a similar manner. We applied a striatum mask (caudate nucleus and putamen taken from the AAL masks implemented in the WFU pickatlas toolbox; Maldjian et al., 2003) and performed the correlation-based MVPA within each searchlight with a radius of 3 voxels (maximum of 123 voxels, and less at the boundaries of the striatum). To claim that a striatal region processes values of social and monetary rewards in a similar manner, within each searchlight, we computed the three following voxel-by-voxel correlations; 1) High Monetary Reward within-condition correlation (i.e., odd vs. even runs), 2) High Social Reward-Self within-condition correlation, and 3) High Monetary Reward vs. High Social Reward-Self between-condition correlation (we took the average of two between-condition correlations). Each correlation was fisher-z transformed and submitted to group level analysis (i.e., one-sample t test [one-tailed]). We looked for regions within the striatum where all three average fisher-transformed correlations are simultaneously significantly positive at $p < 0.05$ level (note that the probability of finding such results by chance is 0.0125% [i.e., $0.05^3 = 0.000125$]) with an extent threshold of 50 contiguous voxels.

RESULTS

Correlation- and classifier-based MVPA in the left putamen and left caudate nucleus ROIs

Since the original univariate GLM analysis identified common activations especially in the left putamen and left caudate nucleus (Izuma et al., 2008), we first focused on these two regions. First, we confirmed the reliability of activation patterns in the two main conditions (High Monetary Reward condition and High Social Reward-Self condition). Each of the two conditions showed a significant within-condition correlation in both the left putamen (both *p*s

< 0.007, Figure 6.2A) and left caudate nucleus (both *p*s < 0.010; Figure 6.2B), indicating that each of these two conditions consistently evoked similar activation patterns across odd and even runs within each of the two ROIs. Interestingly, the average correlation between High Monetary Reward and High Social Reward-Self conditions was significantly positive in the left caudate nucleus (average $r = 0.069$, $t(18) = 2.23$, $p = 0.019$; Figure 6.2B), while it was not significant in the left putamen ($p = 0.43$). To check whether the significant between-condition correlation found in the left caudate nucleus ROI was not due to outliers, we further computed the same correlations after removing outliers (0.23% of the data) based on a Grubbs' test (Grubbs, 1950). The average correlation between High Monetary Reward and High Social Reward-Self conditions was slightly attenuated after removing outliers (average $r = 0.064$), but remained significant ($t(18) = 2.05$, $p = 0.028$).

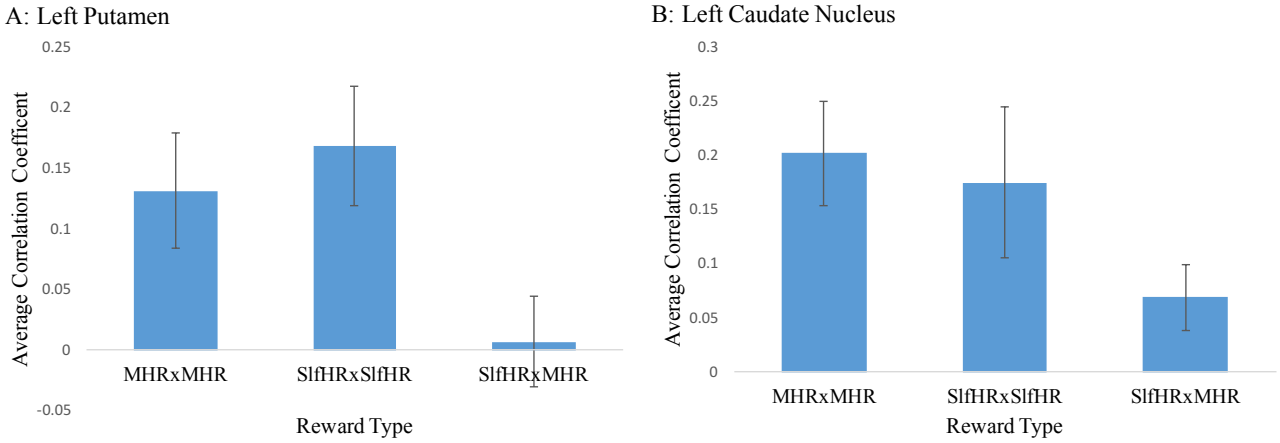


Figure 6.2. Correlation-based MVPA results in the left putamen (**A**) and left caudate nucleus (**B**). MHR: High Monetary Reward, SlfHR: High Social Reward-Self. Error bars denote Standard Error of Mean (SEM).

To check the robustness of the findings in the left caudate nucleus ROI, we further conducted a classifier-based MVPA to test whether a monetary reward classifier can classify social reward and *vice versa*. The result first showed that the monetary reward classifier could distinguish High Monetary Reward vs. No Monetary Reward conditions significantly above the chance level of 50% (average performance = 59.9%, $t(18) = 2.46$, $p = 0.012$). Similarly,

the social reward-self classifier could distinguish High Social Reward-Self and No Social Reward-Self conditions significantly above the chance level (average performance = 56.6%, $t(18) = 2.04$, $p = 0.028$). Importantly, each classifier was generalisable to a different reward type. The monetary reward classifier could distinguish High Social Reward-Self and No Social Reward-Self conditions significantly above the chance level (average performance = 59.9%, $t(18) = 3.75$, $p < 0.001$). Likewise, the social reward classifier could distinguish High Monetary Reward and No Monetary Reward conditions significantly above the chance level (average performance = 55.3%, $t(18) = 3.02$, $p = 0.004$). Furthermore, weight values of the monetary and social reward classifiers were significantly correlated with each other within the left caudate nucleus ROI (average $r = 0.10$, $t(18) = 1.94$, $p = 0.034$). This result indicates that each voxel within the left caudate nucleus similarly contributed to the classification of monetary and social rewards, suggesting shared neural representations between monetary and social rewards within this area.

Exploratory correlation-based MVPA in the four ROIs

We further investigated all possible correlations across nine conditions (3 conditions form the monetary reward experiment and 6 conditions form the social reward experiment) in the putamen and caudate nucleus in both hemispheres (Figure 6.1) to explore detailed representational similarity across all conditions (Figure 6.3A-D). Across all of the four ROIs, for each of the Monetary Reward and Social Reward-Self conditions, the average within-condition correlations were significantly positive (see Figure 6.3E). It should be noted, however, that the average correlations between High Monetary Reward and High Social Reward-Self were significantly positive only in the left caudate nucleus.

Interestingly, we found that the average correlations between High Social Reward-Self and Low Social Reward-Other were all significantly positive across the four ROIs (see Figure 6.3E). As schadenfreude (positive emotion derived from the misfortunate of another individual) is also known to activate the striatum (Cikara et al., 2011; Takahashi et al., 2009), these results may suggest an interesting possibility that two different types of social reward (good reputation toward the self and schadenfreude) share the same neural representations within the human striatum.

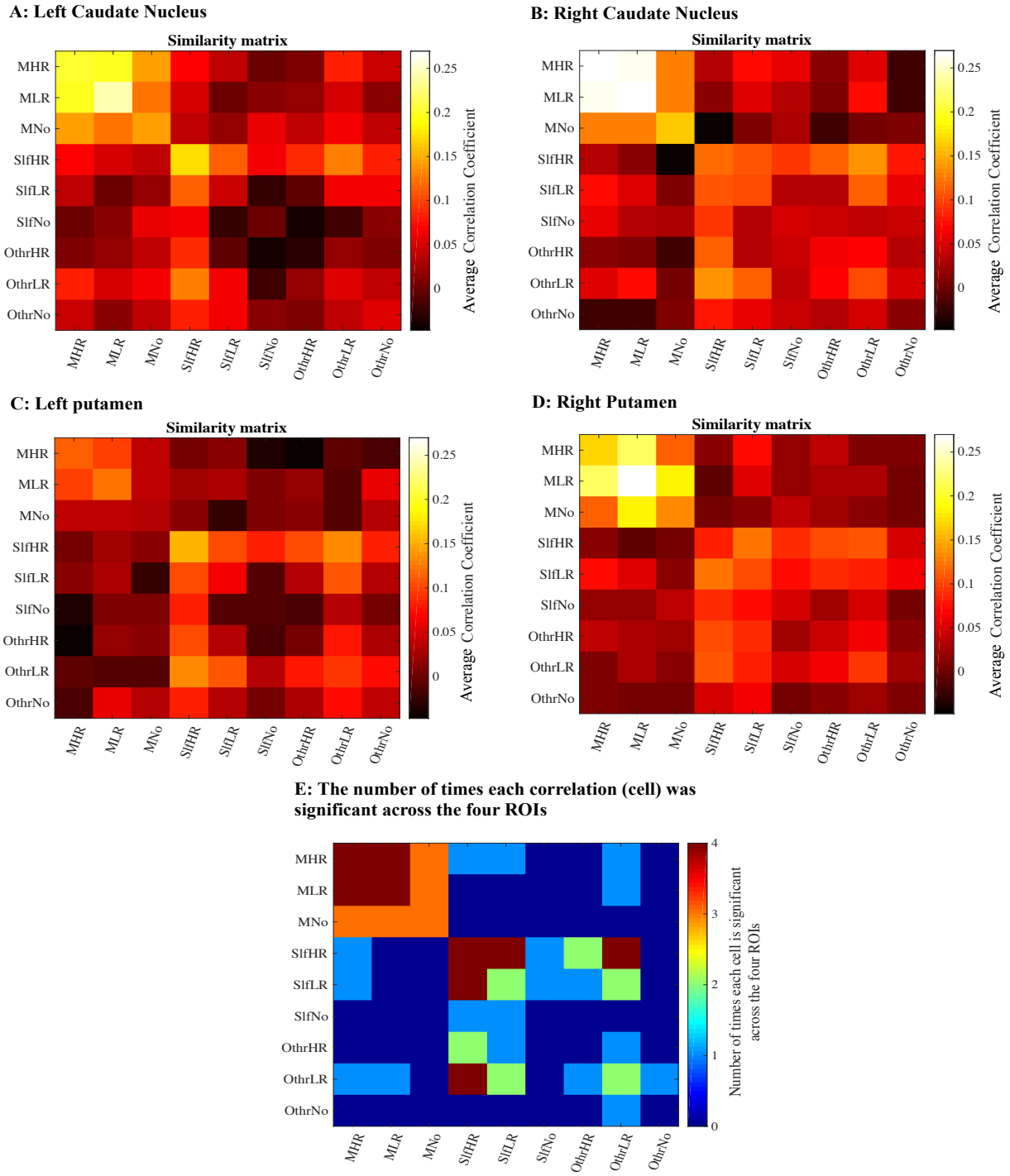


Figure 6.3. Average Correlation similarity matrix in the left caudate nucleus (**A**), the right caudate nucleus (**B**), the left putamen (**C**) and the right putamen (**D**). Each cell represents the average voxel-by-voxel correlation between two conditions across 19 subjects. (**E**) The number of times each average correlation (cell) was significant (based on one-sample t-test) across the four ROIs. MHR: High Monetary Reward condition, MLR: Low Monetary Reward condition, MNo: No Monetary Reward condition, SlfHR: High Social Reward-Self condition, SlfLR: Low Social Reward-Self Condition, SlfNo: No Social Reward-Self condition, OthrHR: High Social Reward-Other condition, OthrLR: Low Social Reward-Other condition, OthrNo: No Social Reward-Other condition.

Exploratory searchlight analysis within the striatum

The searchlight analysis revealed that only in the bilateral ventral striatum (nucleus accumbens; see Figure 6.4) the average correlation between High Monetary Reward and High Social Reward-Self conditions as well as two average within-condition correlations (i.e., odd vs. even runs in the High Monetary Reward condition and odd vs. even runs in the High Social Reward-Self condition) were all significantly positive, suggesting a common neural code for monetary and social rewards in the nucleus accumbens.

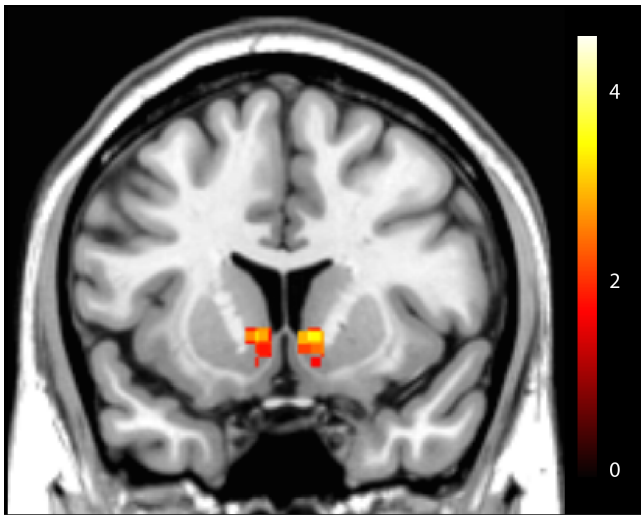


Figure 6.4. Axial slice (y = 12) showing the result of the searchlight analysis. Peak coordinates; left nucleus accumbens (x = -8, y = 16, z = 0, 55 voxels, average r at the peak = 0.089) and right nucleus accumbens (x = 8, y = 16, z = -6, 63 voxels, average r at the peak = 0.109). Colours represent t values based on one-sample t-test testing the strength of the correlation between High Monetary Reward and High Social Reward-Self conditions. Note that the left nucleus accumbens area slightly overlaps (i.e., 9 voxels) with the left caudate ROI (Figure 6.1).

DISCUSSION

Using the correlation and classifier-based MVPA, the present study extends the original study (Izuma et al., 2008) that employed conventional univariate analysis and demonstrated that the left caudate nucleus similarly represents social and monetary rewards. Together with the original finding (Izuma et al., 2008), the left caudate nucleus showed; 1) linear increase in

activation according to reward values of both social and monetary rewards (Izuma et al., 2008), 2) significant voxel-by-voxel correlation between High Monetary Reward and High Social Reward-Self conditions, 3) the Monetary Reward classifier was generalisable to distinguish Social Reward vs. No Social Reward (and *vice versa*), and 4) weight values of Monetary Reward and Social Reward classifiers were significantly correlated with each other, indicating that there is a common neural code for social and monetary rewards in the human striatum. Furthermore, although the left caudate nucleus was the only region that showed a similar representation between two types of reward across the four ROIs (Figure 6.1), the searchlight analysis revealed that the bilateral nucleus accumbens, one of the brain areas most heavily implicated in reward processing (Haber & Knutson, 2010), also represents social and monetary rewards in a similar manner. The results suggest that the same population of neurons within each of these areas encode both abstract social reward as well as physical tangible reward and thus provide support for the “extended common currency schema" (Ruff & Fehr, 2014).

Although significant, the size of the correlations between social and monetary rewards we found in the left caudate ROI and bilateral nucleus accumbens was fairly small (average r = 0.069-0.109; Figure 6.2B and Figure 6.4), suggesting that only a small subset of neurons in this area encodes both social and monetary rewards. This is largely consistent with previous neurophysiological studies. For example, Carelli and Wondolowski (2003) found on a single cell level only 8% of neurons in the nucleus accumbens responded to both juice and drug rewards in rats, and Robinson and Carelli (2008) found that only 15% of nucleus accumbens neurons responded to both juice and ethanol (alcohol) in rats, whereas Bowman, Aigner, and Richmond (1996) found no neurons (0%) in the ventral striatum responded to both juice and drug rewards in monkeys. More recently, Klein and Platt (2013) presented social images (e.g., hindquarters of female monkeys) as reward to monkeys and found that only 6% of striatal neurons encoded information about both juice reward and social images. Thus, although largely

distinct populations of neurons encode different types of reward, there exists a small population of neurons that commonly encode different types of reward in the striatum. The present study further suggests that in the human striatum, there may be the same population of neurons that encode tangible reward and highly abstract social reward of good reputation formed by other people.

Additionally, it may also be noteworthy that we observed similar populations of neurons within the striatum encode information related to receiving high social reward as well as viewing others receiving low social reward. One speculation at this point may suggest similar neural processes occur for social reward and also for the concept of schadenfreude. This falls in line with previous work that reported striatal activation in response to schadenfreude (Cikara et al., 2011; Takahashi et al., 2009) and may suggest a shared neural representation between experiences of schadenfreude and good reputation. Schadenfreude in this sense could suggest a form of reputation management. As social beings flourishing in groups we always have to ensure our place is secure, therefore heightening our own social reputation induces reward, but it may also be that having another more "highly ranked" individual's reputation lowered would still give us the rewarding feeling of amplifying our own group status (in relativity). Aside from this explanation being speculative at this stage, it should also be noted that for Low Social Reward-Other, the result was only significant in two out of four ROIs for the within-condition analysis indicating that activation patterns evoked in this condition are not very consistent. Thus, future research should aim to further dissect this fascinating relationship.

CONCLUSION

In summary, though there have been somewhat discrepant results regarding the encoding of different types of reward in neuroimaging, our results via MVPA indicate that there exists a small population of neurons that commonly encode different types of rewards in the striatum,

and the present study further suggests that in the human striatum, there may be the same population of neurons that encode tangible reward and highly abstract social reward of good reputation formed by other people. This suggests that the brain processes social versus non-social information similarly. Additionally, finding similar neural patterns when participants experience high social reward compared to viewing others receiving low social reward also suggests a potential for similar populations of neurons responsible for processing two different types of social reward (good reputation and schadenfreude). These findings provide an important perspective to some previous research, and help to further illuminate the mechanisms behind social versus non-social cognition.

Chapter 7

A more comprehensive understanding of the neural basis of social information processing

The current thesis can be thought as tackling two key objectives, the first to more fully understand the precise neural and psychological mechanisms involved in processing particularly valent inconsistent social information. This involves the effect social attitudes/orientation has and whether any discrepancies reflect separate psychological mechanisms used as a result of this, and also the more specific role the pMFC has in the detection of social conflict and subsequent behavioural amendments. Secondly, it involves the assessment of whether the human brain includes a specified neural network dedicated exclusively to processing social information. Overall the work presented aids to uncover more fundamental principles involved in the handling and processing of social information. With this, future models can further work towards a more complete understanding of the various levels involved in understanding and responding to social information, aiding in not only the development of scientific understanding, but also paradigms aiming to produce positive behavioural change within social contexts.

Summary of main findings

The first empirical chapter (Chapter 2) aimed to specifically assess whether opposing political orientations could predict similar strength and pattern of activation in response to politically inconsistent material, allowing subsequent insight into the psychological mechanisms regarding political intolerance. Further, this experiment aimed to uncover a more precise neural basis regarding the handling of politically inconsistent information. In this demonstrates two particularly important findings. The first being that no significant difference in univariate activation strength, or multivariate pattern of activation, was seen across participants classed as politically left versus right wing. This suggests the underlying

psychological mechanisms for processing political inconsistent material may be similar across political orientation, supporting the concept that political intolerance can be considered an Ideological Conflict (Brandt et al., 2014), though smaller and imbalanced sample size mean results can only moderately indicate. The second key finding comes from the notion that political information in general (both consistent and inconsistent with participants orientation) tended to possess an activation pattern unique to immoral or generally negative material across key regions associated with processing socially valent information (specifically the left insula, dmPFC, STG, and IFG). Overall results indicate the processing of political material used by the study on a basic level involves typical cognitive processes such as comprehension of syntax (Walenski et al., 2019) and faces (Haxby, Hoffman, & Gobbini, 2000a; Kircher et al., 2000; Mitchell et al., 2005) (evidenced by positive correlations of activation pattern across all conditions in our ROIs), processing the emotional valence of stimuli (indicated by more similar activation pattern of general political, immoral and negative compared to neutral material in the dmPFC and STG), and finally evidence that the left insula, dmPFC, STG, and IFG process general political material significantly more similar than control, negative, *and* immoral material indicates the presence of at least some political specific processes.

The second empirical chapter (Chapter 4) aimed to further dissect the role of the pMFC in processing socially conflicting information, clarifying a role in conflict detection or conflict resolution (i.e. behavioural update/amendment). Using social information designed to elicit cognitive bias, disassociating the level of behavioural update from the level of social conflict presented, this paradigm brought three key findings to light. The first was that participants tended to update beliefs based on new information about others more if it allowed them to see others in a more favourable light, regardless of group membership (i.e. for both in and out group members). Second, a region of the dmPFC tracked the level of conflict across all trials, ascertaining the role of this region in conflict detection. Third, a separate region of the dmPFC

demonstrated increased activation for unfavourable compared to favourable new information about others (again regardless of group membership), demonstrating sensitivity to undesirable versus desirable social outcomes. Since favourable information was more strongly associated with subsequent behavioural update compared to unfavourable information, this suggests the role of the pMFC is more relevant to conflict detection as opposed to conflict resolution.

The final empirical chapter (Chapter 6) aimed to further uncover the notion of a specialised social neural network, dedicated exclusively to the processing and handling of social information in humans. By manipulating the extent of monetary and social reward, the neural circuitry elicited was directly compared using multivariate analysis techniques designed to discriminate the pattern of activation within reward related regions in the brain. In this was found positive correlations between the activation patterns of social vs. monetary reward that indicate neurons encode both social and monetary reward in the ventral striatum (specifically the left caudate nucleus and bilateral nucleus accumbens) similarly, suggesting a common neural code for reward in the human striatum. As this was not also found in the left putamen, and smaller correlation coefficients of activation pattern suggests only a subset of neurons encode both social and non-social reward similarly, further research is needed to establish the extent of a common neural code for social as opposed to non-social reward, and ultimately social versus non-social processes.

Fundamental principles of social information processing

One of the key themes to come from the three empirical chapters as outlined above might be a notion of non-specificity. Ranging from attitudinal orientation, group membership, to social versus non-social, the general consistency of neural responses in associated ROIs is apparent in these particular experiments across the paradigms assessed. For example, behavioural and neural processes were similar across political orientation in the first empirical

chapter, and also in the second across nationality. This indicates that although various psychological models may indicate specifically distinct or specialised underlying processes (for example accounts of right wing intolerance and motivated social cognition by Jost, Glaser, Kruglanski, & Sulloways, 2003), the way we deal with social information, on a neural level at least, may be more generalisable across the specific paradigms assessed than previously alluded to. Overall, this perhaps suggests an absence of particular dedicated systems for externally manufactured subsets of social information processing. However, due to the limitations present in the application of some of these paradigms, for example an imbalance in attitude strength between groups in the first and second empirical chapters, this remains theoretical until further research can replicate the current findings using more well defined and matched social groups.

The third empirical chapter suggests a common neural code regarding social and non-social information amongst subsets of neurons (also echoing a theme of non-specificity), and suggests perhaps some principles from general information processing may be applicable to social models. This is outlined as a possible angle by Adolphs, (2010), who posits that how the brain processes social information may be more computational in nature, being synonymous with information input generally. The current knowledge surrounding social neuroscience could mean regions, in which particular functional heterogeneity is associated, cannot necessarily be applied to exclusively social processes, but instead these regions are simply relevant for social information amongst other general processes. Parkinson and Wheatley (2015) nicely discuss this concept with their neural account of a *repurposed* social brain. This is the idea that brain regions originally purposed for non-social mechanisms, through years of evolutionary pressure to flourish in social interaction and context, are repurposed for specific social cognitions. One example given is the ability to redirect our attention amongst appropriate internal processes, which can support the ability to locate appropriate knowledge relevant to the social context.

Specifically, the authors state one function of the superior parietal lobule (SPL) may have initially evolved to switch attention between external cues but over time also became adapt for internal switching (Shomstein, 2012), supported by similar patterns of activation (indicating shared mechanism) within the SPL for tasks requiring the switching of external and internal attention (Knops, Thirion, Hubbard, Michel, & Dehaene, 2009). Thus, it may be that regions originally specified for various modules of information processing over time became equally specified in social equivalents of information processing. Though this discounts somewhat the notion of a dedicated and specified social brain, our social relevance still make social processes a forefront of brain function.

The current thesis sheds particular light on the role of the pMFC, in particular the dmPFC, a key region across the first two empirical chapters. In this is demonstrated the dmPFC is relevant for the processing of socially inconsistent and conflicting information. What might be interesting is that separate regions of the dmPFC may be involved in the objective as opposed to subjective processing of information. For example and as touched upon earlier, in the second empirical chapter one dmPFC cluster seemed to encode information objectively (the cluster active for general conflict across all trial types, see Chapter 4, Figure 4.4) whereas another dmPFC cluster seemed to be active for personally valent information, such as unfavourable feedback (undesirable outcome) about others (see Chapter 4, Figure 4.6). It may be that within the dmPFC are separate regions that relate to different processing modalities, one for the objective integration of social information, and one that processes information more subjectively, biasing information. Interestingly a study by Kao, Davis, and Gabrieli (2005) found when assessing participants predicted versus actual recollection of images of landscapes, though neural division could primarily be seen for predicted (vmPFC) versus actual (bilateral mid-temporal lobe, left posterior cingulate) outcomes, the dmPFC was linked with both predicted and actual encoding success (although only a trend towards significance was

observed for the association with actual success). The authors suggest from this the dmPFC may relate to both the objective versus subjective valuation of information. Therefore, perhaps relevant functional divisions within the dmPFC are those of objective versus subjective encoding of information, though this account remains speculative until further research examines the dmPFC in this light.

Implications for future work

An important benefit in understanding the neural and psychological basis of social information processing is the ability to aid in positive behavioural change. The more understood about the way social information is integrated, further attempts to isolate predictors that result in negative as opposed to positive appraisals of information can be made. Linking back to work by Parkinson and Wheatley (2015), a pertinent example of this following the account of a repurposed social brain might be in their account of *instrumental* repurposing. In the understanding that older, non-socially-specified neural structures are relevant to process contemporary (often complex) social information, the nature in which information is presented can be manipulated in order to relate to more evolutionary pertinent mechanisms to produce a favourable response. Notably, they argue empathic mechanisms were initially associated (and so are more adapt/responsive) with small, monocultural social groups, explaining the tendency for a proximate, lone individual in trouble to evoke more of an emotional response than the knowledge of, distant, mass poverty (Slovic, 2007). Therefore, the implication of this knowledge means future research could focus interventions that are designed to promote, for example, positive attitude change towards the tolerance of opposing political views in a way specific to tap into evolutionary older mechanisms. As a result, perhaps this would involve the demonstration of an individual's affect surrounding political issues (the deep rooted, personal

reasons one abides by their political stance), rather than presentations of facts that aim to increase tolerance to opposing views.

Another broad goal for social neuroscience, considering the complexity and number of mechanisms involved in processing multifaceted information, might be paradigms designed to incorporate higher ecological validity. This is argued to be an essential step by Schilbach et al., (2013), who claim that paradigms in both psychology and neuroscience studies can unintentionally manufacture effects through the use of third-person-perspective, laboratory stimuli. Rather, they argue aspects of real-time second-person interaction are crucial in understanding the true nature of social information processing, both psychological and neural mechanisms. This is demonstrated in a fMRI experiment that asks participant to imagine being in a social interaction with three other people, whilst a virtual character they see in the scanner directs either socially relevant or arbitrary facial expressions towards the participant (self-directed) or the other imagined characters (other-directed). An increase in neural activity in the vmPFC and amygdala was seen for self-directed facial expressions, whereas other-directed facial expressions were related to varied recruitment of the medial and lateral parietal cortex (Schilbach et al., 2006). This at the very least demonstrates a uniqueness in even subtle forms of social interaction from a second person social experience. Going forward, this approach might map more specific and accurate neural mechanisms involved in social information processing, as opposed to examining neural mechanisms relevant from inferring social interaction via third person perspectives. Not only does more realistic social interaction account for more socially relevant neural correlates, it could also discount other non-relevant processes that may be by-products in more conventional fMRI paradigms. Without ecologically valid paradigms, knowledge specific to social processes cannot be exclusively inferred. This is particularly important if the practical application of knowledge is for positive behavioural change, as outlined above.

Conclusion

Overall, the current doctoral thesis adds the following important ideas to the field of social neuroscience; firstly, political orientation considered from a left-right ideological standpoint does not produce behavioural or neural divergence reflective of previous social accounts that predict distinctive underlying psychological mechanisms. This adds indirect neural support to the notion psychological mechanisms' underlying intolerance stem from contradicting ideology. Secondly, a key neural structure in the processing of inconsistent socio-political/social information in particular is the dmPFC, which demonstrates a key role in representing conflict, and potentially demonstrates functional subdivisions in the objective versus subjective integration of social information. Finally, similar activation patterns between social vs. monetary reward suggest there exists at least a subset of neurons that are responsible for processing information equivalent across social and non-social domains, indicating the neural structures implicated in social information processing may be relevant also across general information processing. With this knowledge, future research can continue to understand the neural accounts of social information processing with more direction. Understanding the processes behind social information processing are essential in the development of the field, which lead to more efficient paradigms to produce positive behavioural change within social contexts. As history and the preceding few years has shown, social conflict via misattribution of social information causes detrimental consequences across society, and so a shift in working towards understanding the mechanisms relevant to the processing of social information in science and importantly society seems worthwhile.

Appendices

Appendix 1a: Political Knowledge, Self report knowledge, Interest, and Discussion measure

Please indicate whether the following statements are true, false, or you don't know:

1. Margaret Thatcher was a Conservative Prime Minister;.....
2. The number of MP's is about 100;.....
3. The longest time allowed between general elections is four years;.....
4. Britain's electoral system is based on proportional representation;.....
5. MPs from different parties are on parliamentary committees;.....
6. Britain has separate elections for the European parliament and the British parliament;.....
7. No-one may stand for parliament unless they pay a deposit;.....

Please now circle the relevant answers on the scales provided to the following questions:

How much, if anything, do you feel you know about 'Politics'?

1	2	3	4
Nothing at all	Not very much	A fair amount	A great deal

How interested would you say you are in "Politics"?

1	2	3	4
Not at all Interested	Not very much	Fairley Interested	Very Interested

How often do you discuss politics or political news with someone else?

1	2	3	4
Never	Rarely	Sometimes	Often

Appendix 1b: Political Intolerance questionnaire

Please read the following statements, and indicate the extent you agree with each one by circling the appropriate numbers on the scale below:

1. I believe that members of Government should not be allowed to organize in order to pass laws demoting (*promoting*) the freedom of movement for European immigrants.

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

2. I think that Thatcherite (*Corbynite*) groups should be allowed to distribute economic policy pamphlets and buttons on local university campuses.

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

3. I think that a group should not be allowed to organize in order to try to decrease (*increase*) the amount of refugees we allow into the country.

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

4. I believe that a group that opposes (*supports*) increases in welfare support should not be allowed to organize in order to influence government policy on welfare support in higher education.

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

5. I believe that a person who supports (*opposes*) the privatization of the NHS should not be allowed to disrupt an MP's meeting.

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

6. I think that a protestor should be allowed to give a speech entitled "Jeremy Corbyn (*Nigel Farage*), Our Generation's Hitler".

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

7. I think that the Britain First Party (*Communist Party of Britain*) should not be allowed to visit university campuses in order to register potential voters.

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

8. I think that protestors who approve (*disapprove*) of introducing more grammar schools in the UK should be allowed to demonstrate in city centres.

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

9. Society shouldn't have to put up with those who have political ideas that are extremely different from the majority.

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

10. It is better to live in an orderly society than to allow people so much freedom that they can become disruptive.

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

11. Free speech is just not worth it if it means that we have to put up with the danger to society of extremist political views.

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

Finally, please indicate on the scales below:

When it comes to economic policy, do you usually consider yourself a liberal, moderate, or conservative?

1	2	3	4	5	6	7
Strongly Liberal			Moderate			Strongly Conservative

When it comes to social policy, do you usually consider yourself a liberal, moderate or conservative?

1	2	3	4	5	6	7
Strongly Liberal			Moderate			Strongly Conservative

(**represented in parenthesis are right wing alternatives*)

Appendix 1c: Empirical Chapter 1 supplementary figure

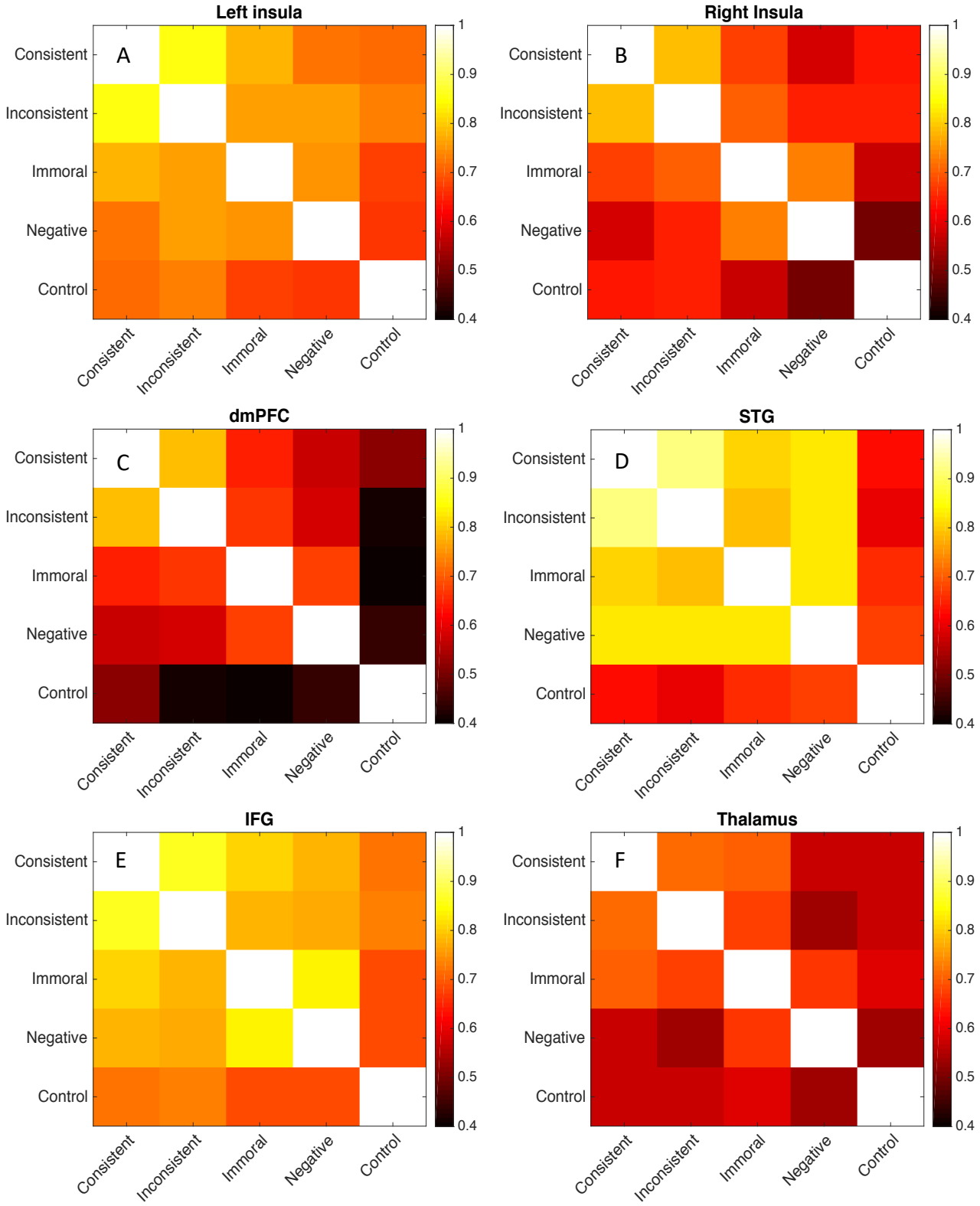


Figure 2.S1. Similarity matrix representing the correlation between all experimental trials in the associated ROIs. Colour bars represent average correlation coefficient.

Appendix 2: Empirical Chapter 2 supplementary materials

Supplementary Methods

Scenarios that do versus don't involve the other-group

To examine any effect involving the associated other-group (i.e. Japanese people acting pro-social/anti-social towards South Korean people, and vice versa) has on the extent participants update information from the first to second estimate, we divided the data into scenarios that do involve the other-group (n=15), and scenarios that don't (n=13). The same GLM set up as reported in the manuscript for both Japanese and South Korean trials was applied within each set of data. Both included predictor variables: 1) First Estimates, 2) Gap (feedback - first estimate, not absolute value), 3) Favourability (dummy coded as favourable = 1 and unfavourable = 0), and 4) Gap × Favourability. All predictors were centred by subtracting the mean value from each score to evade multicollinearity. The dependent variable was Update (second estimate – first estimate).

The effect of "General Favourability": Behavioural Data Analysis

It may be possible that, in addition to the Gap between participant's first estimate and feedback, participants second estimate is influenced by the *General Favourability* of feedback*,* the extent the majority versus minority partake in positive/negative behaviours, regardless of participants expectations. To quantify general favourability of trials, we computed our General Favourability score via the following equation:

If a scenario is positive, General Favourability score = feedback given – 50

If a scenario is negative, General Favourability score = -1 × (feedback given - 50)

Accordingly, this general favourability score takes values between -50 to +50, and a higher score indicates that more people are willing to engage in a positive (pro-social) behaviour or less people are willing to engage in a negative (anti-social) behaviour.

We then entered the following predictor variables into a multiple regression analysis, to assess the degree to which general favourability effects the extent participants updated their second estimates:1) First Estimate 2) General Favourability, 3) Gap (feedback - first estimate, not absolute value), and 4) General Favourability × Gap. All predictors were centred by subtracting the mean value from each score to evade multicollinearity. The dependent variable was Update (second estimate – first estimate). Just like the main regression analyses reported in the manuscript, we ran two separate regression analyses (one for Japanese Trials, and one for South Korean trials).

The effect of "General Favourability": fMRI Data Analysis

We conducted a further GLM on our fMRI data to assess the effect of General Favourability, and the Interaction between General Favourability and Absolute Gap on brain activity. We again used a parametric modulation analysis with a similar set up to our previous analysis to investigate the relationship between trial-by-trial General Favourability scores, Absolute Gap scores, the interaction between General Favourability and Absolute Gap scores, and regional brain activity.

Accordingly, the model included: 1) each trial presentation (duration = total time from onset of initial scenario presentation to onset of feedback presentation), 2) Feedback presentation in Japanese trials (duration = 2 sec), 3) Feedback presentation in Japanese trials modulated by General Favourability (as calculated in the behavioural GLM above), 4) Feedback presentation in Japanese trials modulated by Absolute Gap (duration = 2 sec), 5) Feedback presentation in Japanese trials modulated by General Favourability × Absolute Gap, 6) Feedback presentation in South Korean trials (duration = 2 sec), 7) Feedback presentation in South Korean trials modulated by General Favourability, 8) Feedback presentation in South Korean trials modulated by Absolute Gap (duration = 2 sec), 9)

Feedback presentation in South Korean trials modulated by General Favourability × Absolute Gap, 10) Catch trial presentation (regressor of no interest) (duration = total time of catch trial from initial scenario presentation onset to the end of feedback presentation). This analysis yielded six main contrast images (all Japan and South Korean trials modulated by General Favourability, Absolute Gap, and General Favourability × Absolute Gap) used for second level analysis. Other regressors that were of no interest, such as six motion parameters, the session effect, and high-pass filtering (128 sec) were also included. We again set a whole-brain statistical threshold at $p < 0.001$ voxel wise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple comparisons). Finally, since we incorporate two parametric modulators into our model, we disabled SPMs default implementation of the serial orthogonalization procedure.

The Interaction between Absolute Gap and Update on brain activity: fMRI Data Analysis

We conducted a further GLM on our fMRI data to assess the Interaction between Absolute Gap and Update (corrected for the regression-to-the-mean effect: as used in the second GLM in the manuscript) on brain activity. We again used a parametric modulation analysis with a similar set up to our previous analysis to investigate the relationship between trial-by-trial Absolute Gap scores, Update Scores, the interaction between Absolute Gap and Update scores, and regional brain activity.

Accordingly, the model included: 1) each trial presentation (duration = total time from onset of initial scenario presentation to onset of feedback presentation), 2) Feedback presentation in Japanese Favourable trials (duration = 2 sec), 3) Feedback presentation in Japanese Favourable trials modulated by Absolute Gap (duration = 2 sec), 4) Feedback presentation in Japanese Favourable trials modulated by Update (duration = 2 sec), 5) Feedback presentation in Japanese Favourable trials modulated by Absolute Gap × Update, 6)

Feedback presentation in Japanese Unfavourable trials (duration = 2 sec), 7) Feedback presentation in Japanese Unfavourable trials modulated by Absolute Gap (duration = 2 sec), 8) Feedback presentation in Japanese Unfavourable trials modulated by Update (duration = 2 sec), 9) Feedback presentation in Japanese Favourable trials modulated by Absolute Gap × Update, 10) Feedback presentation in South Korean Favourable trials (duration = 2 sec), 11) Feedback presentation in South Korean Favourable trials modulated by Absolute Gap (duration = 2 sec), 12) Feedback presentation in South Korean Favourable trials modulated by Update (duration = 2 sec), 13) Feedback presentation in South Korean Favourable trials modulated by Absolute Gap × Update, 14) Feedback presentation in South Korean Unfavourable trials (duration = 2 sec), 15) Feedback presentation in South Korean Unfavourable trials modulated by Absolute Gap (duration = 2 sec), 16) Feedback presentation in South Korean Unfavourable trials modulated by Update (duration = 2 sec), 17) Feedback presentation in South Korean Unfavourable trials modulated by Absolute Gap × Update, 18) Catch trial presentation (regressor of no interest) (duration = total time of catch trial from initial scenario presentation onset to the end of feedback presentation). This analysis yielded twelve main contrast images (all of the four conditions modulated by the three parametric regressors; Absolute Gap, Update, and Absolute Gap × Update) used for second level analysis. Other regressors that were of no interest, such as six motion parameters, the session effect, and high-pass filtering (128 sec) were also included. We again set a whole-brain statistical threshold at $p < 0.001$ voxel wise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple comparisons). We again disabled SPMs default implementation of the serial orthogonalization procedure.

Finally, due to a technical fault with the scanner, for one subject, fMRI data after 6 minutes of the first session were not obtained. Accordingly, the fMRI data analysis included 144 images for the first session (it should have been 214 images). In this session, the subject

still continued the task without being scanned for approximately 3 minutes so that our behavioural data analysis included all trials. Due to the small number of trials in the condition for this particular analysis with an interaction variable, we excluded the first session data for this subject (for all other GLMs, the analysis of this subject's fMRI data included the 144 images from the first session).

The Interaction between Update and Favourability on brain activity: fMRI Data Analysis

We conducted a further GLM on our fMRI data to assess the Interaction between Update (corrected for the regression-to-the-mean effect: as used in the second GLM in the manuscript) and Favourability on brain activity. We again used a parametric modulation analysis with a similar set up to our previous analysis to investigate the relationship between trial-by-trial Update Scores, Favourability, and the interaction between Update scores and Favourability, and regional brain activity.

Accordingly, the model included: 1) each trial presentation (duration = total time from onset of initial scenario presentation to onset of feedback presentation), 2) Feedback presentation in Japanese trials (duration = 2 sec), 3) Feedback presentation in Japanese trials modulated by Update (duration = 2 sec), 4) Feedback presentation in Japanese trials modulated by Favourability (duration = 2 sec), 5) Feedback presentation in Japanese trials modulated by Update × Favourability, 6) Feedback presentation in South Korean trials (duration = 2 sec), 7) Feedback presentation in South Korean trials modulated by Update (duration = 2 sec), 8) Feedback presentation in South Korean trials modulated by Favourability (duration = 2 sec), 9) Feedback presentation in South Korean trials modulated by Update × Favourability, 10) Catch trial presentation (regressor of no interest) (duration = total time of catch trial from initial scenario presentation onset to the end of feedback presentation). This analysis yielded six main contrast images (both Japanese and South

Korean conditions modulated by the three parametric regressors; Update, Favourability, and Update × Favourability) used for second level analysis. Other regressors that were of no interest, such as six motion parameters, the session effect, and high-pass filtering (128 sec) were also included. We again set a whole-brain statistical threshold at $p < 0.001$ voxel wise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple comparisons). We again disabled SPMs default implementation of the serial orthogonalization procedure.

Supplementary Results

The effect of involving the other-group: Behavioural Results

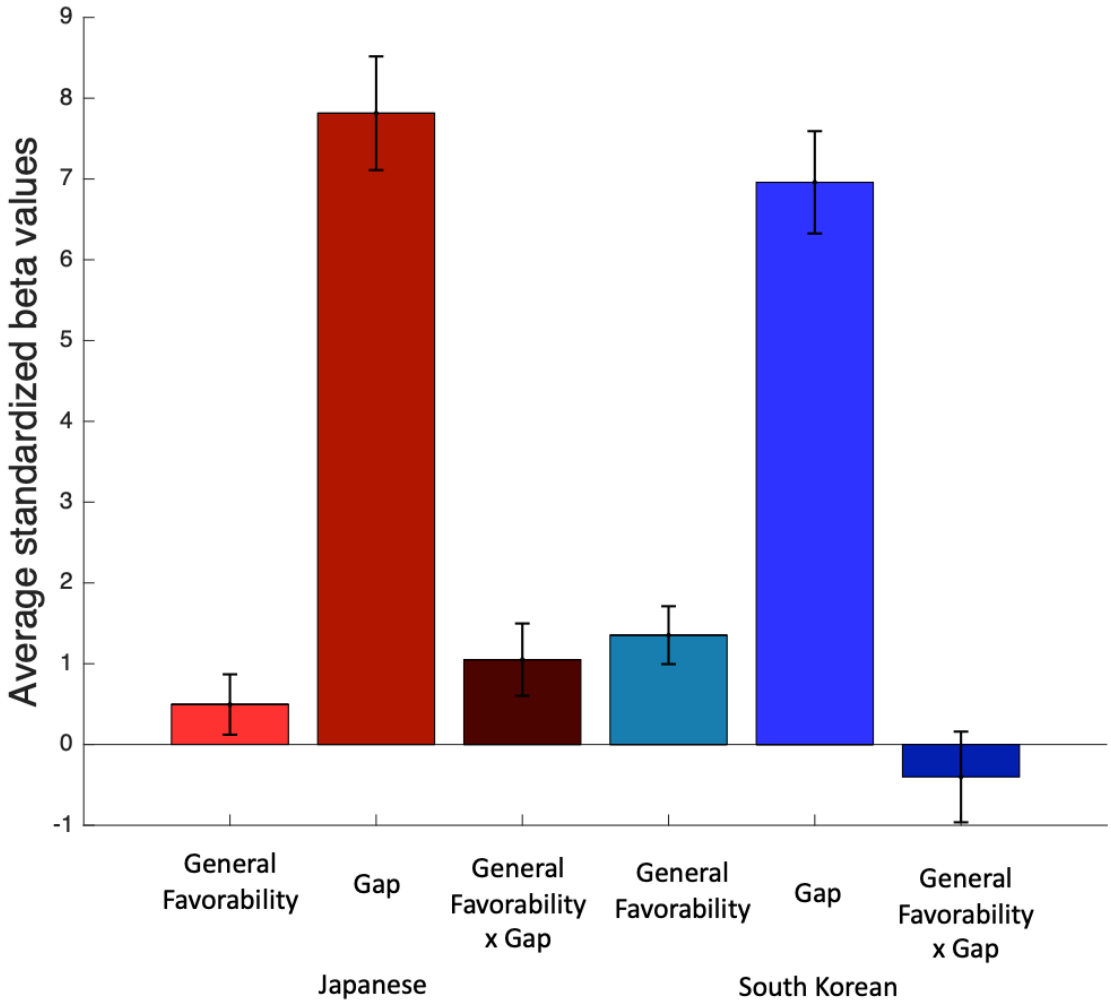
We see no significant different at the group level for statements that involve vs. don't involve the other-group for Gap in Japanese (p= 0.62) or South Korean trials (p= 0.12), First estimate in Japanese (p = 0.10) or South Korean trials (p= 0.50), and Favourability for Japanese (p= 0.33) or South Korean Trials (p= 0.81). However, though we don't see a significant difference for Japanese trials that do or don't involve the other-group for the interaction Gap × Favourability (p= 0.83), we do see a significant difference for South Korean Trials ($t(27)$= -2.19, p= 0.04). This demonstrates that participants were more likely to update information in response to our Favourability bias significantly more when scenarios *don't* involve the other-group (but note that p values were not corrected for multiple comparisons). Overall, there was no strong difference between the two types of scenarios, and the result suggests the general pro-social nature of our scenarios, rather than any action towards the other-group specifically, is what primarily drives our effects.

The effect of "General Favourability": Behavioural Results

We again found a significant effect of *Gap* (feedback - first estimate) for Japanese ($t(27) = 11.10$ $p < 0.001$) and South Korean trials ($t(27) = 10.99$ $p < 0.001$), meaning that

participants updated their scores *more* from the first to the second rating the larger the gap was between their first rating and the feedback they were presented with. We found a significant effect of General Favourability for South Korean trials ($t(27) = 3.78$ $p < 0.001$), meaning trials that were generally more favourable regardless of participants expectation or initial estimate were updated more from participants first to second rating, though we don't see this effect for Japanese trials ($p = 0.20$). We observed a significant interaction effect of General Favourability × Gap for Japanese trials ($t(27)= 2.35$, $p = 0.03$) meaning that participants updated their scores significantly more in response to generally favourable feedback when Gap was larger. However, we don't observe the same interaction for South Korean trials ($p = 0.48$).

To further investigate any variation between Japanese and South Korean trials, variation in strength of effect between General Favourability, Gap, and General Favourability × Gap, and any interaction present, we conducted a 2 × 3 within-subjects ANOVA [Nationality (Japanese Trials vs. South Korean Trials) × Predictor (General Favourability vs. Gap vs. General Favourability × Gap)] on the standardised beta values. We found a significant main effect for Predictor $F(2,54)= 69.43$, $p < 0.001$, but no significant main effect for Nationality ($p = 0.92$). Further, we see a significant interaction between Predictor × Nationality $F(2,54)= 4.60$, $p = 0.01$. A series of (Bonferroni-Holm) corrected paired t-tests demonstrated the effect of General Favourability was significantly stronger for South Korean compared to Japanese trials $t(27)= -2.73$, $p = 0.03$. Alternatively, the effect of General Favourability × Gap was significantly stronger for Japanese compared to South Korean trials $t(27)= 2.40$, $p = 0.047$. There was no significant difference between Japanese versus South Korean trials for Gap ($p = 0.20$) (see Supplementary Figure 4.S1).



Supplementary Figure 4.S1. Bars represent average standardised beta values for Japan (red shaded) and South Korea (blue shaded). Error bars denote SEM.

Overall, consistent with the results of the original regression model (Table 4.1), this highlights the strong effect of Gap. The results of this new regression analyses further suggest that participants are only influenced by General Favourability for Japanese trials when this is more unexpected (i.e. General Favourability significantly interacts with Gap, but we observe no main effect). On the other hand, for South Korean trials participants are more broadly influenced by the General Favourability of trials, and the effect of General Favourability was not modulated by any prior expectations (i.e. significant main effect of General Favourability, but no interaction with Gap).

The effect of "General Favourability": fMRI Results

When broadly depicting regions related to General Favourability across all trials regardless of nationality (i.e., Japan and South Korea conditions combined), we didn't find any significant cluster for this contrast, and this was also the case for Japanese and South Korean trials separately. The regions related to Absolute Gap across trials were similar to our initial analysis examining just the modulation of Absolute Gap (see Supplementary Table 4.S1 for all results), for example a significant cluster in the dmPFC. In assessment of regions related to the interaction between General Favourability and Absolute Gap across all trials, no voxels passed our threshold. This was also the case for Japanese and South Korean trials separately.

Therefore, we conclude that although the General Favourability of trials has an impact on participants behavioural Update (further supporting our conclusion on the favourability bias shown with our participants, i.e. the tendency to update information that allows them to see others more positive/favourable), it seems this effect has little impact on brain activity as measured in the current paradigm, and doesn't significantly interact with our Gap/Absolute Gap variable on a behavioural or neural level.

The Interaction between Absolute Gap and Update on brain activity: fMRI Results

In order to first replicate regions related to the Absolute Gap across all trials regardless of nationality or favourability, we find the results are similar to our initial analysis examining just the modulation of Absolute Gap (see Table 4.2 in the main manuscript), for example a significant cluster in the pMFC. Furthermore, examination of brain regions *negatively* correlated with Absolute Gap again revealed significant activation within the ventral striatum (see Supplementary Table 4.S2). To again assess the regions related to the Update of all trials, we replicated our initial result and didn't find any significant cluster for this contrast. In

assessment of the regions related to the interaction between Absolute Gap and Update across all trials, we didn't find any significant activation.

Therefore, we found the same independent contribution (as the two main fMRI GLMs in the manuscript assessing Absolute Gap and Update separately) from the regressors combined in this model but failed to see any effect from the Interaction between the two on brain activity.

The Interaction between Update and Favourability on brain activity: fMRI Data Analysis

In order to first view regions related to the Update of trials regardless of nationality (i.e., Japan and South Korea conditions combined), we replicated our initial result and didn't find any significant cluster for this contrast. This was also the case for Japanese and South Korean trials separately. To assess the regions related to Favourability across all trials, we show a significant cluster in the left middle frontal gyrus (see Supplementary Table 4.S3). In assessment of the interaction between Update and Favourability, we didn't find any significant activation. Therefore, from the regressors combined in this model we fail to see any effect from the Interaction between Update and Favourability on brain activity.

Overlap in activation with Izuma and Adolphs (2013)

Although the dmPFC regions sensitive to Absolute Gap (Figure 4.4) in the present study are slightly lateralised within the dmPFC compared to Izuma and Adolphs (2013) previous study, there was considerable overlap between them (270 voxels). There is also overlap in the PCC (263 voxels) and left IFG (269 voxels), suggesting the pattern of whole brain activations is similar. We likewise found overlap between the dmPFC region identified in Izuma and Adolphs (2013) and the dmPFC region especially sensitive to Absolute Gap for Unfavourable trials compared to Favourable trials (Figure 4.6; 108 voxels). This might

suggest that the dmPFC is generally sensitive to the difference between one's rating and the reality or group opinion, but activity in a part of the dmPFC (such as the green region depicted in Figure 4.6) is modulated by what an individual *hopes* the reality or group opinion to be (and the dmPFC cluster reported in Izuma and Adolphs (2013) included both of these regions).

Supplementary Tables

Supplementary Table 4.S1. Brain regions correlated with Absolute Gap, General favourability, and General Favourability x Absolute Gap

Location	BA	MNI coordinate				Cluster size
		x	y	z	Z	
Areas positively correlated with Absolute Gap						
dmPFC	8	-8	26	60	5.53	2225
Left superior temporal gyrus (STG)	22	-54	-36	2	5.74	2510
left inferior frontal gyrus (IFG)	47	-44	32	-10	5.70	
Right medial frontal-orbital gyrus	18	2	16	-24	5.45	405
Right STG	44	44	16	36	5.30	491
Right IFG	47	46	32	-10	5.05	407
Right Precuneus	23	6	-54	34	4.72	863
Areas positively correlated with General Favourability						
*no significant clusters						
Areas positively correlated with General Favourability x Absolute Gap						
*no significant clusters						

BA, Brodmann area. Statistics are based on a set threshold of height $p < 0.001$ (uncorrected), and cluster $p < 0.05$ (FWE). Areas in grey italics represent significant peak (FWE) sub-clusters/different regions within larger clusters.

Supplementary Table 4.S2. Brain regions correlated with Absolute Gap, Update, and Absolute Gap × Update

Location	BA	MNI coordinate				Cluster size
		x	y	z	Z	
Areas positively correlated with Absolute Gap						
dmPFC	9	-8	52	36	4.14	402
Areas negatively correlated with Absolute Gap						
Left postcentral gyrus	40	-58	-32	48	4.55	258
Left nucleus accumbens	25	-14	6	-12	4.50	843
Areas positively correlated with Update						
*no significant clusters						
Areas positively correlated with Update × Absolute Gap						
*no significant clusters						

BA, Brodmann area. Statistics are based on a set threshold of height $p < 0.001$ (uncorrected), and cluster $p < 0.05$ (FWE). Areas in grey italics represent significant peak (FWE) sub-clusters/different regions within larger clusters.

Supplementary Table 4.S3. Brain regions correlated with Update, Favourability, and Update × Favourability.

Location	BA	MNI coordinate				Cluster size
		x	y	z	Z	
Areas positively correlated with Update						
*no significant clusters						
Areas positively correlated with Favourability						
Left middle frontal gyrus	6	-42	0	42	4.45	443
Areas positively correlated with Update × Favourability						
*no significant clusters						

BA, Brodmann area. Statistics are based on a set threshold of height $p < 0.001$ (uncorrected), and cluster $p < 0.05$ (FWE). Areas in grey italics represent significant peak (FWE) sub-clusters/different regions within larger clusters.

Supplementary Table 4.S4. Brain regions correlated with Absolute Gap separately for each condition.

Location	BA	MNI coordinate				Cluster size
		x	y	z	Z	
Japanese Favourable: Areas positively correlated with Absolute Gap						
*no significant clusters						
Japanese Unfavourable: Areas positively correlated with Absolute Gap						
*no significant clusters						
South Korean Favourable: Areas positively correlated with Absolute Gap						
Left superior frontal gyrus (SFG)	8	-14	36	56	4.28	756
Left Precuneus	7	0	-56	36	4.45	200
South Korean Unfavourable: Areas positively correlated with Absolute Gap						
*no significant clusters						

BA, Brodmann area. Statistics are based on a set threshold of height $p < 0.001$ (uncorrected), and cluster $p < 0.05$ (FWE).

Supplementary Table 4.S5. Brain regions associated with Absolute Gap for Favourable>Unfavourable trials, and Unfavourable>Favourable trials

Location	BA	MNI coordinate			Z	Cluster size
		x	y	z		
Favourable>Unfavourable: Areas positively correlated with Absolute Gap						
*no significant clusters						
Unfavourable>Favourable: Areas positively correlated with Absolute Gap						
dmPFC	8	6	38	48	4.35	238
Left IFG	48	-48	18	22	5.38	1137
Left middle frontal gyrus (MFG)	44	-52	16	42	5.01	
Right MFG	6	40	8	58	4.69	607
Right middle occipital gyrus (MOG)	19	32	-66	32	3.92	235

BA, Brodmann area. Statistics are based on a set threshold of height $p < 0.001$ (uncorrected), and cluster $p < 0.05$ (FWE). Areas in grey italics represent significant peak (FWE) sub-clusters/different regions within larger clusters.

Full list of Scenarios

South Korean Positive *(Japanese alternative simply switched the nationality involved from Japanese to South Korean)*

1. Would you be willing to help up a Japanese person in public if they fell over?
2. Would you be willing to share a post on Facebook that compliments Japanese culture?
3. If you accidently ripped your friends favourite scarf, would you tell them truthfully what had happened and offer to buy them a new one?
4. Do you believe that showing equal respect to everyone, no matter their class or ethnicity, is more important to educate into society than academic education?
5. Do you believe it is justified for a university lecturer to tell students that it's always wrong for governments to not attempt to solve tension between Korea and Japan?
6. Do you believe it is a good lesson for a mother to teach her children that if any of their friends at school don't have any lunch, they should always share their lunch?
7. Would you be willing to accept a Facebook friend request off a Japanese person you met on a trip?
8. If you found your friend posting offensive tweets about Japanese people, would you try to stop it?
9. If you saw a bird caught in some litter at the side of the road, would you pull over to help it?
10. Would you remain calm and polite despite a train passenger being extremely rude towards you due to a misunderstanding over a train seat?
11. Do you believe it is a positive thing that a husband chooses to stay home from work to take care of his sick child while his wife had an important meeting?
12. Would you be willing to regularly give some food to homeless people?

13. If someone at work was wanting to get promoted so they could earn more money to provide for their family, would you be willing to help them?
14. Would you allow your child to play with a Japanese child?
15. Do you believe a college class that's aim is to inform students on reducing discrimination towards the Japanese in general society is a positive thing?
16. Do you believe it is a positive thing for males and females to earn equal pay?
17. Do you believe it is a positive thing that Korean companies hire Japanese people?
18. Would you ever attempt to save a Japanese person from getting mugged if you could?
19. If there was a big earthquake in Japan, would you be willing to donate foreign aid?
20. Would you believe it to be a positive thing if a government meeting between South Korea and Japan occurred with the intention to build better relations between the countries?
21. If a Japanese student lost their house keys, would you provide them with shelter for the night?
22. Do you think it is a positive thing for a South Korean government to try to improve gender equality in a society?
23. Would you be willing to give your seat to an elderly or a pregnant woman even if you are very tired?
24. If you found a lost wallet on a street, would you be willing to bring it to a police station even if no one saw you picking it up?
25. Would you be willing to wait to hold the door for an old person?
26. If an old person holding a large suitcase is having trouble getting up the stairs, would you help this person go up the stairs?
27. If natural disasters such as earthquakes occurred in other areas, would you decide to go volunteer to help their recovery?

28. If a Japanese student asks you to teach them Korean, would you help him/her?

South Korean Negative *(Japanese alternative simply switched the nationality involved from Japanese to South Korean)*

1. If you saw racist material towards Japanese people on social media, would you feel positive about it?
2. Do you believe it is acceptable that when intoxicated at a party people sometimes vandalise property?
3. Do you think it's ok to search for and watch YouTube videos that make fun of Japanese people?
4. If a person accidently drops some money, would you ever be willing to take it rather than inform them?
5. If a person loses their temper badly because they received bad service in a restaurant, do you believe this behaviour is justified?
6. Under extreme circumstances, would you ever be willing to launch a nuclear missile towards Japan?
7. Would you ever start a riot in Japan if you thought you were receiving unfair treatment due to you being Korean?
8. Do you believe it is justified for a Korean person to start a fight with a Japanese person because they insulted their country?
9. Would you ever be willing to let a colleague take the blame for your mistake in order to avoid getting into trouble?
10. Would you ever be willing to take credit for someone else's work who has recently left the company in order to make yourself look good?
11. Would you ever be willing to drop litter on the street because there wasn't a bin nearby?

12. Do you believe it is a positive lesson for a teacher to encourage students to betray their friends if it meant them getting ahead in their career?
13. Do you believe it is justified that a young boy chooses to go out with his friends rather than staying home to help his grandmother whilst she was feeling unwell?
14. Would you ever be willing to use someone else's milk in the work fridge without asking who it belongs to?
15. Do you believe it is justified that a grandfather doesn't allow his grandchildren to watch a film because a Japanese actor stars in it?
16. If you ever broke your mother's favourite ornament, would you then lie that you did not break it when asked about it?
17. Would you laugh at a joke your friend told you that is rude to Japanese people?
18. Do you believe it is justified for a Korean tourist to write a blog while visiting Japan that is very offensive and paints Japanese people in a negative way?
19. Would you ever deliberately not clear your tray from a fast food restaurant and leave it for someone else to do because you were too tired?
20. If you were able to, would you ever be willing to push in front of someone in a que because you were in a rush?
21. If a Japanese person stopped you to ask for directions, would you pretend you didn't know the way, even though you did?
22. Would you ever avoid sitting next to a Japanese person on a train?
23. In a South Korean city, a restaurant posts a sign saying "Korean Only" Do you believe this is justified?
24. If you received more (cash) change than you should, would you walk away without reporting it?
25. Do you think it is justified to text while walking on a busy street if you are in a rush?

26. Would you ever purposely drive too close to someone if you were in a rush to hurry them up?
27. Would you ever jaywalk across a street if no one is watching you?
28. Do you think it is justified to break a promise with your friend if circumstances change?

References

Adolphs, R. (2003). Cognitive neuroscience: Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience*, *4*(3), 165–178. https://doi.org/10.1038/nrn1056

Adolphs, R. (2010). Conceptual challenges and directions for social neuroscience. *Neuron*, *65*(6), 752–767. https://doi.org/10.1016/j.neuron.2010.03.006

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1995). Fear and the human amygdala. *The Journal of Neuroscience*, *15*(9), 5879–5891. https://doi.org/10.1523/jneurosci.15-09-05879.1995

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., Sanford, R. N., Aron, B. R., Levinson, M. H., & Morrow, W. R. (1950). *The authoritarian personality*. New York: Harper and Row.

Allport, G. W. (1954). *The nature of prejudice*. Reading, Mass.: Addison-Wesly.

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*, *99*(3), 436.

Amodio, D. M., Jost, J. T., Master, S. L., & Yee, C. M. (2007). Neurocognitive correlates of liberalism and conservatism. *Nature Neuroscience*, *10*(10), 1246–1247. https://doi.org/10.1038/nn1979

Apperly, I. (2010). *Mindreaders: The Cognitive Basis of "Theory of Mind."*(1st Editio). London: Psychology Press. https://doi.org/10.4324/9780203833926

Asch, S. E. (1952). Group forces in the modification and distortion of judgments. In *Social psychology.* (pp. 450–501). Englewood Cliffs: Prentice-Hall, Inc. https://doi.org/10.1037/10025-016

Barch, D. M., Braver, T. S., Akbudak, E., Conturo, T., Ollinger, J., & Snyder, A. (2001). Anterior Cingulate Cortex and Response Conflict: Effects of Response Modality and Processing Domain. *Cerebral Cortex*, *11*(9), 837–848. https://doi.org/10.1093/cercor/11.9.837

Bench, C. J., Frith, C. D., Grasby, P. M., Friston, K. J., Paulesu, E., Frackowiakt, R. S. J., & Dolantt, R. J. (1991). Investigations of the functional anatomy of attention using the stroop test. *Neuropsychologia*, *31*(9), 907–922.

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The Cross-Category Effect. *Psychological Science*, *18*(8), 706–712. https://doi.org/10.1111/j.1467-9280.2007.01964.x

Berntsen, D. (2002). Tunnel memories for autobiographical events: Central details are remembered more frequently from shocking than from happy experiences. *Memory & Cognition*, *30*(7), 1010–1020.

Bigler, E. D., Ozonoff, S., Krasny, L., Lu, J., Provencal, S. L., Mcmahon, W., & Lainhart, J. E. (2007). Superior Temporal Gyrus , Language Function, and Autism. *Developmental Neuropsychology*, *31*(2), 217–238.

Blum, B. (2018, March). The Lifespan of a Lie. https://doi.org/10.1348/014466605X81720

Bond, M. H., & Smith, P. B. (1996). Cross-cultural social and organizational psychology. *Annu. Rev. Psychol*, *47*, 205–240.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict Monitoring and Cognitive Control. *Cohen & Servan-Schreiber*, *108*(3), 624–652. https://doi.org/10.1037//0033-295X.I08.3.624

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, *8*(12), 539–546.

Botvinick, M., Nystrom, L., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, 402, 179–181.

Bowman, E. M., Aigner, T. G., & Richmond, B. J. (1996). Neural signals in the monkey ventral striatum related to motivation for juice and cocaine rewards. *Journal of Neurophysiology*, *75*(3), 1061–1073. https://doi.org/10.1152/jn.1996.75.3.1061

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The Ideological-Conflict Hypothesis. *Current Directions in Psychological Science*, *23*(1), 27–34. https://doi.org/10.1177/0963721413510932

Brothers, L. (1990). The social brain : A project for integrating primate behaviour and neurophysiology in a new domain. In *Concepts in Neuroscience* (Vol. 1, pp. 27–51). Retrieved from https://ci.nii.ac.jp/naid/20000695363/

Bush, G., & Shin, L. M. (2006). The Multi-Source Interference Task: an fMRI task that reliably activates the cingulo-frontal-parietal cognitive/attention network. *Nature Protocols*, *1*(1), 308–313. https://doi.org/10.1038/nprot.2006.48

Cacioppo, J. T., Berntson, G. G., & Decety, J. (2010). Social neuroscience and its relatioship to social psychology. *Social Cognition*, *28*(6), 675–685.

Cacioppo, J. T., & Decety, J. (2012). *An Introduction to Social Neuroscience*. The Oxford Handbook of Social Neuroscience (pp. 3–8). New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195342161.013.0001

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the opinion of others affects our valuation of objects. *Current Biology*, *20*(13), 1165–1170.

Caprara, G. V., Schwartz, S., Capanna, C., Vecchione, M., & Barbaranelli, C. (2006). Personality and politics: Values, traits, and political choice. *Political Psychology*, *27*(1), 1–28.

Carelli, R. M., & Wondolowski, J. (2003). Selective encoding of cocaine versus natural rewards by nucleus accumbens neurons is not related to chronic drug exposure. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *23*(35), 11214–11223.

Carleton, R. N., Collimore, K. C., & Asmundson, G. J. G. (2010). "It's not just the judgements—It's that I don't know": Intolerance of uncertainty as a predictor of social anxiety. *Journal of Anxiety Disorders*, *24*(2), 189–195. https://doi.org/10.1016/J.JANXDIS.2009.10.007

Carney, D. R., Jost, J. T., Gosling, S. D., & Potter, J. (2008). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political Psychology*, *29*(6), 807–840.

Chapman, H. A., & Anderson, A. K. (2013). Things rank and gross in nature: a review and synthesis of moral disgust. *Psychological Bulletin*, *139*(2), 300.

Chen, M. K., & Risen, J. L. (2010). How choice affects and reflects preferences: Revisiting the free-choice paradigm. *Journal of Personality and Social Psychology*, *99*(4), 573–594. https://doi.org/10.1037/a0020217

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them: social identity shapes neural responses to intergroup competition and harm. *Psychological Science*, *22*(3), 306–313. https://doi.org/10.1177/0956797610397667

Clark, L. A., Watson, D., & Mineka, S. (1994). Temperament, personality, and the mood and anxiety disorders. Special Issue: Personality and psychopathology. *Journal of Abnormal Psychology*, *103*(1), 103–116.

Crawford, J. T., & Pilanski, J. M. (2014). Political Intolerance. *Source: Political Psychology*, *35*(6), 841–851. https://doi.org/10.1111/j.l

Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin*, *115*(1), 74–101. https://doi.org/10.1037/0033-2909.115.1.74

Crisp, R. J., & Meleady, R. (2012). Adapting to a multicultural future. *Science*, *336*(6083), 853–855.

Crisp, R. J., & Turner, R. N. (2011). Cognitive adaptation to the experience of social and cultural diversity. *Psychological Bulletin*, *137*(2), 242.

Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7*(2), 189.

Cunningham, W. A., Raye, C. L., & Johnson, M. K. (2004). Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of Cognitive Neuroscience*, *16*(10), 1–3.

Cunningham, W. A., & Zelazo, P. D. (2007). Attitudes and evaluations: A social cognitive

neuroscience perspective. *Trends in Cognitive Sciences*, *11*(3), 97–104.

Damasio, A. R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L. B., Parvizi, J., & Hichwa, R. D. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience*, *3*(10), 1049.

Dávid-Barrett, T., & Dunbar, R. I. M. (2013). Processing power limits social group size: Computational evidence for the cognitive costs of sociality. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1765). https://doi.org/10.1098/rspb.2013.1151

De Quervain, D. J. F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*(5688), 1254–1258. https://doi.org/10.1126/science.1100735

De Zubicaray, G. I., Andrew, C., Zelaya, F. O., Williams, S. C. R., & Dumanoir, C. (2000). Motor response suppression and the prepotent tendency to respond: a parametric fMRI study. *Neuropsychologia*, *38*(9), 1280–1291.

Delgado, M. R. (2007). Reward-related responses in the human striatum. In *Annals of the New York Academy of Sciences* (Vol. 1104, pp. 70–88). https://doi.org/10.1196/annals.1390.002

Dodge, K. A. (2014). A social information processing model of social competence in children. In *Cognitive perspectives on children's social and behavioral development* (pp. 85–134). Psychology Press.

Duchaine, B., & Yovel, G. (2015). A Revised Neural Framework for Face Processing. *Annual Review of Vision Science*, *1*(1), 393–416. https://doi.org/10.1146/annurev-vision-082114-035518

Dunbar, R. (1998). The social brain hypothesis. *Evolutionary Anthropology*, *6*, 178–190.

Dunbar, R. I. M., & Shultz, S. (2007a). Understanding primate brain evolution. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 362, pp. 649–658). https://doi.org/10.1098/rstb.2006.2001

Dunbar, R. I. M., & Shultz, S. (2007b, September 7). Evolution in the social brain. *Science*. American Association for the Advancement of Science. https://doi.org/10.1126/science.1145463

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138.

Eisenberger, N. I. (2012). The pain of social disconnection: examining the shared neural underpinnings of physical and social pain. https://doi.org/10.1038/nrn3231

Emonds, G., Declerck, C. H., Boone, C., Vandervliet, E. J. M., & Parizel, P. M. (2012). The cognitive demands on cooperation in social dilemmas: An fMRI study. *Social Neuroscience*, *7*(5), 494–509. https://doi.org/10.1080/17470919.2012.655426

Falk, E. B., Cascio, C. N., O'Donnell, M. B., Carp, J., Tinney Jr, F. J., Bingham, C. R., … Simons-Morton, B. G. (2014). Neural responses to exclusion predict susceptibility to social influence. *Journal of Adolescent Health*, *54*(5), S22–S31.

Fan, J., Flombaum, J. I., McCandliss, B. D., Thomas, K. M., & Posner, M. I. (2002). Cognitive and brain consequences of conflict. *NeuroImage*, *18*(1), 42–57. https://doi.org/10.1006/nimg.2002.1319

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*, *11*(10), 419–427. https://doi.org/10.5167/uzh-2518

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*(6), 939–946.

Ferrari, C., Vecchi, T., Todorov, A., & Cattaneo, Z. (2016). Interfering with activity in the dorsomedial prefrontal cortex via TMS affects social impressions updating. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(4), 626–634.

Festinger, L. (1962). *A theory of cognitive dissonance* (Vol. 2). Stanford university press.

Fischer, P., Greitemeyer, T., & Frey, D. (2008). Self-Regulation and Selective Exposure: The Impact of Depleted Self-Regulation Resources on Confirmatory Information Processing. *Journal of Personality and Social Psychology*, *94*(3), 382–395. https://doi.org/10.1037/0022-3514.94.3.382

Floden, D., & Stuss, D. T. (2006). Inhibitory control is slowed in patients with right superior medial frontal damage. *Journal of Cognitive Neuroscience*, *18*(11), 1843–1849.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (2002). Action representation and the inferior parietal lobule. *Common Mechanisms in Perception and Action Attention and Performance Vol XIX*, *19*, 247–266.

Garavan, H., Ross, T. J., & Stein, E. A. (1999). Right hemispheric dominance of inhibitory control: an event-related functional MRI study. *Proceedings of the National Academy of Sciences*, *96*(14), 8301–8306.

Garrett, N., González-Garzón, A., Foulkes, L., Levita, L., & Sharot, T. (2018). Updating Beliefs Under Perceived Threat. *The Journal of Neuroscience*, *38*(36), 7901–7911. https://doi.org/10.2139/ssrn.3155415

Gilbert, D. T., & Hixon, J. G. (1991). The Trouble of Thinking Activation and Application of Stereotypic Beliefs. *Journal of Personality and Social Psychology*, *60*(4), 509–517.

Glassner, B., & Tajfel, H. (2006). Social Identity and Intergroup Relations. *Contemporary Sociology*, *14*(4), 520. https://doi.org/10.2307/2069233

Golby, A. J., Gabrieli, J. D. E., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience*, *4*(8), 845–850. Retrieved from http://neurosci.nature.com

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400.

Greening, S. G., Finger, E. C., & Mitchell, D. G. V. (2011). Parsing decision making processes in prefrontal cortex: Response inhibition, overcoming learned avoidance, and reversal learning. *NeuroImage*, *54*, 1432–1441. https://doi.org/10.1016/j.neuroimage.2010.09.017

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197.

Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics*, *21*(1), 27–58. https://doi.org/10.1214/aoms/1177729885

Haber, S. N., & Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *35*(1), 4–26. https://doi.org/10.1038/npp.2009.129

Hackel, L. M., Zaki, J., & Van Bavel, J. J. (2017). Social identity shapes social valuation: evidence from prosocial behavior and vicarious reward. *Social Cognitive and Affective Neuroscience*, *12*(8), 1219–1228. https://doi.org/10.1093/scan/nsx045

Hamann, S., & Mao, H. (2002). Positive and negative emotional verbal stimuli elicit activity in the left amygdala. *Neuroreport*, *13*(1), 15–19. https://doi.org/10.1097/00001756-200201210-00008

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations. *Science*, *316*(5831), 1622–1625. https://doi.org/10.1126/science.1140738

Harmon-Jones, E., Gerdjikov, T., & Harmon-Jones, C. (2008). The effect of induced compliance on relative left frontal cortical activity: A test of the action-based model of dissonance. *European Journal of Social Psychology*, *38*(1), 35–45.

Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, *14*(1), 40–48. https://doi.org/10.1016/j.tics.2009.10.011

Hastorf, A. H., & Cantril, H. (1951). Case reports they saw a game: a case study. In *princeton review* (pp. 129–134).

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000a). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Haxby, J. V, Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430.

Haxby, J. V, Hoffman, E. A., & Gobbini, M. I. (2000b). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of Psychology*, *21*(1), 107–112.

Hewstone, M. (1990). The 'ultimate attribution error'? A review of the literature on intergroup causal attribution. *European Journal of Social Psychology*, *20*(4), 311–335. https://doi.org/10.1002/ejsp.2420200404

Hills, P. J., & Lewis, M. B. (2011). Rapid communication: The own-age face recognition bias in children and adults. *Quarterly Journal of Experimental Psychology*, *64*(1), 17–23. https://doi.org/10.1080/17470218.2010.537926

Hofmann, W., Schmeichel, B. J., & Baddeley, A. D. (2012, March 1). Executive functions and self-regulation. *Trends in Cognitive Sciences*. Elsevier Current Trends. https://doi.org/10.1016/j.tics.2012.01.006

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679.

Huang, Y., Zhen, S., & Yu, R. (2019). Distinct neural patterns underlying ingroup and outgroup conformity. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(11), 4758–4759. https://doi.org/10.1073/pnas.1819421116

Hughes, B. L., Zaki, J., & Ambady, N. (2017). Motivation alters impression formation and related neural systems. *Social Cognitive and Affective Neuroscience*, *12*(1), 49–60. https://doi.org/10.1093/scan/nsw147

Hutchinson, D. (2014). "Continually Reminded of Their Inferior Position": Social Dominance, Implicit Bias, Criminality, and Race. *Washington University Journal of Law & Policy*, *46*. Retrieved from http://openscholarship.wustl.edu/law_journal_law_policy/vol46/iss1/8

Izuma, K. (2013). The neural basis of social influence and attitude change. *Current Opinion in Neurobiology*, *23*(3), 456–462.

Izuma, K., & Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron*, *78*(3), 563–573.

Izuma, K., Aoki, R., Shibata, K., & Nakahara, K. (2019). Neural signals in amygdala predict implicit prejudice toward an ethnic outgroup. *NeuroImage*, *189*, 341–352. https://doi.org/10.1016/j.neuroimage.2019.01.019

Izuma, K., Matsumoto, M., Murayama, K., Samejima, K., Sadato, N., & Matsumoto, K. (2010). Neural correlates of cognitive dissonance and choice-induced preference change. *Proceedings of the National Academy of Sciences*, 201011879.

Izuma, K., & Murayama, K. (2013). Choice-Induced Preference Change in the Free-Choice Paradigm: A Critical Methodological Review. *Frontiers in Psychology*, *4*, 41. https://doi.org/10.3389/fpsyg.2013.00041

Izuma, K., & Murayama, K. (2019). The neural basis of cognitive dissonance. In E. Harmon-Jones (Ed.), Cognitive dissonance: Re-examining a Pivotal Theory in Psychology (2nd edition). Washington, DC: American Psychological Association.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of Social and Monetary Rewards in the Human Striatum. *Neuron*, *58*(2), 284–294. https://doi.org/10.1016/J.NEURON.2008.03.020

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: its structure, functions, and elective affinities. *Annual Review of Psychology*, *60*, 307–337. https://doi.org/10.1146/annurev.psych.60.110707.163600

Jost, J. T., Glaser, J., Sulloway, F. J., & Kruglanski, A. W. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375. https://doi.org/10.4324/9781315175867

Jost, J. T., Nam, H. H., Amodio, D. M., & Van Bavel, J. J. (2014). Political neuroscience: The beginning of a beautiful friendship. *Political Psychology*, *35*(SUPPL.1), 3–42. https://doi.org/10.1111/pops.12162

Kanai, R., Feilden, T., Firth, C., & Rees, G. (2011). Political Orientations Are Correlated with Brain Structure in Young Adults. *Current Biology*, *21*(8), 677–680. https://doi.org/10.1016/J.CUB.2011.03.017

Kao, Y.-C., Davis, E. S., & Gabrieli, J. D. E. (2005). Neural correlates of actual and predicted memory formation. *Nature Neuroscience*, *8*(12), 1776–1783. https://doi.org/10.1038/nn1595

Kaplan, J. T., Freedman, J., & Iacoboni, M. (2007). Us versus them: Political attitudes and party affiliation influence neural response to faces of presidential candidates. *Neuropsychologia*, *45*, 55–64. https://doi.org/10.1016/j.neuropsychologia.2006.04.024

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports*, *6*(1), 39589. https://doi.org/10.1038/srep39589

Kaplan, J. T., Man, K., & Greening, S. G. (2015). Multivariate cross-classification: applying machine learning techniques to characterize abstraction in neural representations. *Frontiers in Human Neuroscience*, *9*, 151. https://doi.org/10.3389/fnhum.2015.00151

Kemmelmeier, M. (2008). Is there a relationship between political orientation and cognitive ability? A test of three hypotheses in two studies. *Personality and Individual*

Differences, *45*(8), 767–772. https://doi.org/10.1016/j.paid.2008.08.003

Kensinger, E. A. (2007). Negative emotion enhances memory accuracy: Behavioral and neuroimaging evidence. *Current Directions in Psychological Science*, *16*(4), 213–218.

Kerns, G. J., Cohen, J. D., Macdonald, W. A., Cho, Y. R., Stenger, V. A., & Carter, C. S. (2004). Anterior Cingulate Conflict Monitoring and Adjustments in Control. *Science*, *423*(2), 32. https://doi.org/10.1126/science.1091611

Kircher, T. T. ., Senior, C., Phillips, M. L., Benson, P. J., Bullmore, E. T., Brammer, M., … David, A. S. (2000). Towards a functional neuroanatomy of self processing: effects of faces and words. *Cognitive Brain Research*, *10*(1–2), 133–144. https://doi.org/10.1016/S0926-6410(00)00036-7

Kirk, A., & Dunford, D. (2017). EU referendum: How the results compare to the UK's educated, old and immigrant populations. *The Telegraph*. Retrieved from https://www.telegraph.co.uk/politics/2016/06/24/eu-referendum-how-the-results-compare-to-the-uks-educated-old-an/

Kitayama, S., Snibbe, A. C., Markus, H. R., Suzuki, T., & Snibbe, X. A. C. (2004). Is there any "free" choice? Self and Dissonance in Two Cultures. *Psychological Science*, *15*(8), 527–533.

Klein, J. T., & Platt, M. L. (2013). Social Information Signaling by Neurons in Primate Striatum. *Current Biology*, *23*(8), 691–696. https://doi.org/10.1016/J.CUB.2013.03.022

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140–151.

Klumpp, H., Fitzgerald, D. A., & Phan, K. L. (2013). Neural predictors and mechanisms of cognitive behavioral therapy on threat processing in social anxiety disorder. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *45*, 83–91. https://doi.org/10.1016/J.PNPBP.2013.05.004

Knobloch-Westerwick, S., Johnson, B. K., & Westerwick, A. (2015). Confirmation Bias in Online Searches: Impacts of Selective Exposure Before an Election on Political Attitude Strength and Shifts. *Journal of Computer-Mediated Communication*, *20*(2), 171–187. https://doi.org/10.1111/jcc4.12105

Knobloch-Westerwick, S., Mothes, C., Johnson, B. K., Westerwick, A., & Donsbach, W. (2015). Political online information searching in Germany and the United States: Confirmation bias, source credibility, and attitude impacts. *Journal of Communication*, *65*(3), 489–511.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*(5800), 829–832.

Knops, A., Thirion, B., Hubbard, E. M., Michel, V., & Dehaene, S. (2009). Recruitment of an area involved in eye movements during mental arithmetic. *Science (New York, N.Y.)*,

324(5934), 1583–1585. https://doi.org/10.1126/science.1171599

Knutson, K. M., Wood, J. N., Spampinato, M. V, & Grafman, J. (2006). Politics on the brain: an FMRI investigation. *Social Neuroscience*, *1*(1), 25–40. https://doi.org/10.1080/17470910600670603

Konishi, S., Nakajima, K., Uchida, I., Kikyo, H., Kameyama, M., & Miyashita, Y. (1999). Common inhibitory mechanism in human inferior prefrontal cortex revealed by event-related functional MRI. *Brain*, *122*(5), 981–991.

Korn, C. W., Fan, Y., Zhang, K., Wang, C., Han, S., & Heekeren, H. R. (2014). Cultural influences on social feedback processing of character traits. *Frontiers in Human Neuroscience*, *8*, 192. https://doi.org/10.3389/fnhum.2014.00192

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively Biased Processing of Self-Relevant Social Feedback. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Krishnan, A., Woo, C.-W., Chang, L. J., Ruzic, L., Gu, X., Ló pez-Solà, M., … Wager, T. D. (2016). Somatic and vicarious pain are represented by dissociable multivariate brain patterns. https://doi.org/10.7554/eLife.15166.001

Kross, E., Berman, M. G., Mischel, W., Smith, E. E., & Wager, T. D. (2011). Social rejection shares somatosensory representations with physical pain. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(15), 6270–6275. https://doi.org/10.1073/pnas.1102693108

Kruglanski, A. W., Jasko, K., Milyavsky, M., Chernikova, M., Webber, D., Pierro, A., … Webber, D. (2018). Cognitive Consistency Theory in Social Psychology : A Paradigm Reconsidered. *Psychological Inquiry*, *29*(2), 45–59. https://doi.org/10.1080/1047840X.2018.1480619

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., … Turner, R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(12), 5675–5679.

Larcinese, V. (2007). Does political knowledge increase turnout? Evidence from the 1997 British general election. *Public Choice*, *131*(3–4), 387–411.

Lee, C.-S. (1985). *Japan and Korea. The Political Dimension. Pacific Affairs* (Vol. 59). Hoover press. https://doi.org/10.2307/2758349

Lee, H. Y., Kléber, M., Hari, L., Brault, V., Suter, U., Taketo, M. M., … Sommer, L. (2004). Instructive Role of Wnt/β-Catenin in Sensory Fate Specification in Neural Crest Stem Cells. *Science*, *303*(5660), 1020–1023. https://doi.org/10.1126/science.1091611

Lemerise, E. A., & Arsenio, W. F. (2000). An integrated model of emotion processing. *Child Development*, *71*(1), 107–118.

Leung, H. C., Skudlarski, P., Gatenby, J. C., Peterson, B. S., & Gore, J. C. (2000). An Event-related Functional MRI Study of the Stroop Color Word Interference Task. *Cerebral Cortex*, *10*(6), 552–560. https://doi.org/10.1093/cercor/10.6.552

Levy, B. R., & Banaji, M. R. (2002). Implicit Ageism. In M. Cambridge (Ed.), *Ageism: Stereotyping and Prejudice against Older Persons*. MIT Press. https://doi.org/10.7551/mitpress/10679.003.0006

Lin, L. C., Qu, Y., & Telzer, E. H. (2018). Intergroup social influence on emotion processing in the brain. *Proceedings of the National Academy of Sciences*, *115*(42), 10630–10635. https://doi.org/10.1073/pnas.1802111115

Lindquist, K. A., Satpute, A. B., Wager, T. D., Weber, J., & Barrett, L. F. (2015). The brain basis of positive and negative affect: evidence from a meta-analysis of the human neuroimaging literature. *Cerebral Cortex*, *26*(5), 1910–1922.

Llinás, R. R. (1977). *The Biology of the Brain: From Neurons to Networks: Readings from Scientific American*. Freeman.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098.

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, *97*(8), 4398–4403. https://doi.org/10.1073/pnas.070039597

Maguire, E. A., Spiers, H. J., Good, C. D., Hartley, T., Frackowiak, R. S. J., & Burgess, N. (2003). Navigation Expertise and the Human Hippocampus: A Structural Brain Imaging Analysis. *Hippocampus*, *13*, 208–217. https://doi.org/10.1002/hipo.10087

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, *19*(3), 1233–1239. https://doi.org/10.1016/S1053-8119(03)00169-1

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, *13*(4), 330–334. https://doi.org/10.1037/h0028434

Mansouri, F. A., Egner, T., & Buckley, M. J. (2017). Monitoring demands for executive control: shared functions between human and nonhuman primates. *Trends in Neurosciences*, *40*(1), 15–27.

Martin, A., & Weisberg, J. (2003). Neural foundations for understanding social and mechanical concepts. *Cognitive Neuropsychology*, *20*(3–6), 575–587. https://doi.org/10.1080/02643290342000005

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure and Function*, *214*(5–6), 655–667. https://doi.org/10.1007/s00429-010-0262-0

Milgram, S. (1963). Behavioral Study of Obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371-378. https://doi.org/https://psycnet.apa.org/doi/10.1037/h0040525

Mitchell, J. P., Neil Macrae, C., & Banaji, M. R. (2005). Forming impressions of people versus inanimate objects: Social-cognitive processing in the medial prefrontal cortex. *NeuroImage*, *26*(1), 251–257. https://doi.org/10.1016/J.NEUROIMAGE.2005.01.031

Molenberghs, P. (2013). The neuroscience of in-group bias. *Neuroscience and Biobehavioral Reviews*, *37*, 1530–1536. https://doi.org/10.1016/j.neubiorev.2013.06.002

Molenberghs, P., Halász, V., Mattingley, J. B., Vanman, E. J., & Cunnington, R. (2013). Seeing is believing: Neural mechanisms of action-perception are biased by team membership. *Human Brain Mapping*, *34*(9), 2055–2068. https://doi.org/10.1002/hbm.22044

Molenberghs, P., & Louis, W. R. (2018). Insights From fMRI Studies Into Ingroup Bias. *Frontiers in Psychology*, *9*, 1868. https://doi.org/10.3389/fpsyg.2018.01868

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(42), 15623–15628. https://doi.org/10.1073/pnas.0604475103

Montague, P. R., & Lohrenz, T. (2007). To Detect and Correct: Norm Violations and Their Enforcement. *Neuron*, *56*(1), 14–18. https://doi.org/10.1016/J.NEURON.2007.09.020

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109

Negative views of Russia on the Rise: Global Poll. (2014). Retrieved from https://globescan.com/wp-content/uploads/2014/06/2014_country_rating_poll_bbc_globescan.pdf

Nickerson, R. S. (1998). *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*. *Review of General Psychology* (Vol. 2). Retrieved from https://pdfs.semanticscholar.org/70c9/3e5e38a8176590f69c0491fd63ab2a9e67c4.pdf

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Ogawa, S., & Lee, T. M. (1990). Magnetic resonance imaging of blood vessels at high fields: in vivo and in vitro measurements and image simulation. *Magnetic Resonance in Medicine*, *16*(1), 9–18.

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of*

Sciences of the United States of America, *87*(24), 9868–9872.

Ogawa, S., Lee, T. M., Nayak, A. S., & Glynn, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, *14*(1), 68–78.

Oxley, D. R., Smith, K. B., Alford, J. R., Hibbing, M. V, Miller, J. L., Scalora, M., … Hibbing, J. R. (2008). Political attitudes vary with physiological traits. *Science*, *321*(5896), 1667–1670.

Parkinson, C., & Wheatley, T. (2015). The repurposed social brain. *Trends in Cognitive Sciences*. *19*(3), 133–141. https://doi.org/10.1016/j.tics.2015.01.003

Peelen, M. V, & Downing, P. E. (2007). Using multi-voxel pattern analysis of fMRI data to interpret overlapping functional activations. *Trends in Cognitive Sciences*, *11*(1), 4.

Peyron, R., Laurent, B., & Garcia-Larrea, L. (2000). Functional imaging of brain responses to pain. A review and meta-analysis (2000). *Neurophysiologie Clinique/Clinical Neurophysiology*, *30*(5), 263–288.

Pine, A., Sadeh, N., Ben-Yakov, A., Dudai, Y., & Mendelsohn, A. (2018). Knowledge acquisition is governed by striatal prediction errors. *Nature Communications*, *9*(1), 1673. https://doi.org/10.1038/s41467-018-03992-5

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, *72*(5), 692–697.

Prater, K. E., Hosanagar, A., Klumpp, H., Angstadt, M., & Phan, K. L. (2013). Aberrant amygdala-frontal cortex connectivity during perception of fearful faces and at rest in generalized social anxiety disorder. *Depression and Anxiety*, *30*(3), 234–241. https://doi.org/10.1002/da.22014

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(04), 515. https://doi.org/10.1017/S0140525X00076512

Premkumar, P., Ettinger, U., Inchley-Mort, S., Sumich, A., Williams, S. C. R., Kuipers, E., & Kumari, V. (2012). Neural processing of social rejection: the role of schizotypal personality traits. *Human Brain Mapping*, *33*(3), 695–706.

Reeck, C., Ames, D. R., & Ochsner, K. N. (2016). The Social Regulation of Emotion: An Integrative, Cross-Disciplinary Model. *Trends in Cognitive Sciences*, *20*(1), 47–63. https://doi.org/10.1016/j.tics.2015.09.003

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron*, *35*(2), 395–405.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport*, *15*(16), 2243–2539.

Robinson, D. L., & Carelli, R. M. (2008). Distinct subsets of nucleus accumbens neurons encode operant responding for ethanol versus water. *European Journal of Neuroscience*, *28*(9), 1887–1894.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549.

Rushworth, M. F., Buckley, M. J., Behrens, T. E., Walton, M. E., & Bannerman, D. M. (2007, April 1). Functional organization of the medial frontal cortex. *Current Opinion in Neurobiology*. Elsevier Current Trends. https://doi.org/10.1016/j.conb.2007.03.001

Rushworth, M. F. S., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2004.07.009

Saalmann, Y. B., & Kastner, S. (2011). Cognitive and Perceptual Functions of the Visual Thalamus. *Neuron*, *71*(2), 209–223. https://doi.org/10.1016/J.NEURON.2011.06.027

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin*, *141*(1), 213.

Saxe, R. (2010). Theory of Mind (Neural Basis). In *Encyclopedia of Consciousness* (pp. 401–409). https://doi.org/10.1016/B978-012373873-8.00078-5

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *The Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R., & Vogeley, K. (2006). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia*, *44*, 718–730. https://doi.org/10.1016/j.neuropsychologia.2005.07.017

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, *10*(3), 272–283.

Seehausen, M., Kazzer, P., Bajbouj, M., Heekeren, H. R., Jacobs, A. M., Klann-Delius, G., … Prehn, K. (2014). Talking about social conflict in the MRI scanner: neural correlates of being empathized with. *NeuroImage*, *84*, 951–961.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Sherrill, K. R., Chrastil, E. R., Aselcioglu, I., Hasselmo, M. E., & Stern, C. E. (2018). Structural Differences in Hippocampal and Entorhinal Gray Matter Volume Support Individual Differences in First Person Navigational Ability. *Neuroscience*, *380*, 123–131. https://doi.org/10.1016/j.neuroscience.2018.04.006

Shomstein, S. (2012). Cognitive functions of the posterior parietal cortex: top-down and bottom-up attentional control. *Frontiers in Integrative Neuroscience*, *6*, 38. https://doi.org/10.3389/fnint.2012.00038

Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*(5661), 1157–1162.

Singer, T., Seymour, B., O'doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*(7075), 466.

Slovic, P. (2007). "If I look at the mass I will never act ": Psychic numbing and genocide. *Judgment & Decision Making*, *2*(2), 79–95.

Smith, A. (2013). Civic Engagement in the Digital Age. Retrieved May 29, 2019, from www.pewresearch.org

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law*, *7*(1), 36–97. https://doi.org/10.1037/1076-8971.7.1.36

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Sunstein, C. R., Bobadilla-Suarez, S., Lazzaro, S. C., & Sharot, T. (2016). How people update beliefs about climate change: Good news and bad news. *Cornell L. Rev.*, *102*, 1431.

Swick, D., Ashley, V., & Turken, U. (2008). Left inferior frontal gyrus is critical for response inhibition. *BMC Neuroscience*, *9*(1), 102.

Tajfel, H. (2003). Social Psychology of Intergroup Relations. *Annual Review of Psychology*, *33*(1), 1–39. https://doi.org/10.1146/annurev.ps.33.020182.000245

Tajfel, Henri. (1969). Cognitive aspects of prejudice. *Journal of Biosocial Science*, *1*(S1), 173–191. https://doi.org/10.1017/S0021932000023336

Takahashi, H., Kato, M., Matsuura, M., Mobbs, D., Suhara, T., & Okubo, Y. (2009). When your gain is my pain and your pain is my gain: neural correlates of envy and schadenfreude. *Science*, *323*(5916), 937–939.

Takahashi, H., Yahata, N., Koeda, M., Matsuda, T., Asai, K., & Okubo, Y. (2004). Brain activation associated with evaluative processes of guilt and embarrassment: An fMRI study. *NeuroImage*, *23*(3), 967–974. https://doi.org/10.1016/j.neuroimage.2004.07.054

Talati, A., & Hirsch, J. (2005). Functional Specialization within the Medial Frontal Gyrus for Perceptual Go/No-Go Decisions Based on "What," "When," and "Where" Related Information: An fMRI Study. *Journal of Cognitive Neuroscience*, *17*(7), 981–993. https://doi.org/10.1162/0898929054475226

Turner, C. (2018). Eight in ten British university lecturers are “Left-wing”, survey finds, pp. 2–5. Retrieved from https://www.telegraph.co.uk/education/2017/03/02/eight-ten-british-university-lecturers-left-wing-survey-finds/

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The Neural Substrates of In-Group Bias A Functional Magnetic Resonance Imaging Investigation. *Psycholgical Science*, *19*(11), 1131–1139.

van Prooijen, J.-W., & Krouwel, A. P. M. (2017). Extreme political beliefs predict dogmatic intolerance. *Social Psychological and Personality Science*, *8*(3), 292–300.

Van Veen, V., Krug, M. K., Schooler, J. W., & Carter, C. S. (2009). Neural activity predicts attitude change in cognitive dissonance. *Nature Neuroscience*, *12*(11), 1469.

Verfaellie, M., & Heilman, K. M. (1987). Response preparation and response inhibition after lesions of the medial frontal lobe. *Archives of Neurology*, *44*(12), 1265–1271.

Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: A voxel-based meta-analysis. *Journal of Cognitive Neuroscience*, *22*(12), 2864–2885.

Wake, S.J., & Izuma, K. (2017). A common neural code for social and monetary rewards in the human striatum. *Social Cognitive and Affective Neuroscience*, *12*(10). https://doi.org/10.1093/scan/nsx092

Walenski, M., Europa, E., Caplan, D., & Thompson, C. K. (2019). Neural networks for sentence comprehension and production: An ALE-based meta-analysis of neuroimaging studies. *Human Brain Mapping*, *40*(8), 2275–2304. https://doi.org/10.1002/hbm.24523

Welborn, B. L., & Lieberman, M. D. (2018). Neuropsychologia Disconfirmation modulates the neural correlates of the false consensus e ff ect : A parametric modulation approach. *Neuropsychologia*, *121*, 1–10. https://doi.org/10.1016/j.neuropsychologia.2018.09.018

Westen, D., Blagov, P., Harenski, K., & Kilts, C. (2006). An fMRI Study of Motivated Reasoning: Partisan Political Reasoning in the US Presidential Election. *Journal of Cognitive Neuroscience*, *18*(11), 1947–1958.

Wetherell, G. A., Brandt, M. J., & Reyna, C. (2013). Discrimination Across the Ideological Divide: The Role of Value Violations and Abstract Values in Discrimination by Liberals and Conservatives. *Social Psychological and Personality Science*, *4*(6), 658–667. https://doi.org/10.1177/1948550613476096

Wilson, T. D., & Bar-Anan, Y. (2008). Psychology: The unseen mind. *Science*, *321*(5892), 1046–1047. https://doi.org/10.1126/science.1163029

Woo, C.-W., Koban, L., Kross, E., Lindquist, M. A., Banich, M. T., Ruzic, L., … Wager, T. D. (2014). Separate neural representations for physical pain and social rejection. *Nature Communications*, *5*(1), 5380. https://doi.org/10.1038/ncomms6380

Wright, D. B., & Sladden, B. (2003). An own gender bias and the importance of hair in face

recognition. *Acta Psychologica*, *114*(1), 101–114. https://doi.org/10.1016/S0001-6918(03)00052-0

Wright, J.. C., McWhite, C. B., & Grandjean, P. T. (2014). The cognitive mechanisms of intolerance. In *Oxford Studies in Experimental Philosophy* (Vol 1). Oxford University press.

Wright, Jennifer Cole, Cullum, J., & Schwab, N. (2008). The cognitive and affective dimensions of moral conviction: Implications for attitudinal and behavioral measures of interpersonal tolerance. *Personality and Social Psychology Bulletin*, *34*(11), 1461–1476. https://doi.org/10.1177/0146167208322557

Wu, H., Luo, Y., & Feng, C. (2016). Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *71*, 101–111.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational Substrates of Norms and Their Violations during Social Exchange. *Journal of Neuroscience*, *33*(3), 1099–1108. https://doi.org/10.1523/jneurosci.1642-12.2013

Yoo, S.-S., Choi, B.-G., Juh, R., Pae, C.-U., & Lee, C.-U. (2005). Head motion analysis during cognitive fMRI examination: application in patients with schizophrenia. *Neuroscience Research*, *53*(1), 84–90.

Zamboni, G., Gozzi, M., Krueger, F., Duhamel, J.-R., Sirigu, A., & Grafman, J. (2009). Individualism, conservatism, and radicalism as criteria for processing political beliefs: a parametric fMRI study. *Social Neuroscience*, *4*(5), 367–383.

Zimbardo, P. G., & White, G. (1972). The Stanford Prison Experiment: A simulation study of the psychology of imprissityonment conducted August 1971 at Stanford University [On-line slide show]. *Zimbardo Comprisonexp*.