University of Southampton

Faculty of Environmental and Life Sciences

Psychology

The Neural Representation of The Self

by

Marie Levorsen

ORCID ID 0000-0003-1158-1659

Thesis for the degree of Doctor of Philosophy

October 2025

University of Southampton

Abstract

Faculty of Environmental and Life Sciences

Psychology

Doctor of Philosophy

Neural Representation of The Self

by

Marie Levorsen

The main aim of my PhD thesis was to add knowledge about how the self (i.e., the self-concept) is represented in the brain. I completed three empirical projects that I present across three empirical chapters. In the first empirical chapter (Chapter 2), I report two fMRI experiments with a searchlight RSA approach to investigate where and how the self is represented in the brain. I found that the self is represented in the mPFC in terms of self-importance, but not self-descriptiveness. In the second empirical chapter (Chapter 3), I conducted a behavioural experiment. I used an evaluative priming paradigm to test the psychological meaning of the associative links that connect self-related concepts to the self in an associate network model of the self. I hypothesised that the associative links represent self-importance. The hypothesis was disconfirmed. In the third empirical chapter (Chapter 4), I conducted an fMRI experiment with an MVPA and RSA searchlight approach. I tested similarities and differences in neural patterns of activation for the self-reference task compared to three other tasks known to activate the mPFC, that is, the other-reference task, an autobiographical memory task, and an introspection task. I found that some patterns of activation in the mPFC are shared across the self-reference task and the other three tasks, whereas other patterns of activation are specific to the self. Taken together, the findings contribute to understanding how information about the self is represented in the brain and open up new research directions.

Table of Contents

Table of Tables

Table of Figures

Table of Figures

Research Thesis: Declaration of Authorship

Print name: Marie Levorsen

Title of thesis: Neural Representation of The Self

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

Levorsen, M., Aoki, R., Matsumoto, K., Sedikides, C., & Izuma, K. (2023). The self-concept is represented in the medial prefrontal cortex in terms of self-importance. *Journal of Neuroscience*, *43*(20), 3675-3686. https://doi.org/10.1523/JNEUROSCI.2178-22.2023 (Chapter 2)

Levorsen, M., Aoki, R., Sedikides, C., & Izuma, K. (2025). Decomposing cognitive processes in the mPFC during self-thinking. *Journal of Neuroscience*, *45*(22), e2378242025. https://doi.org/10.1523/JNEUROSCI.2378-24.2025 (Chapter 4)

Signature: .....Date:.....

Acknowledgements

To my supervisor, Keise Izuma, thank you for being so patient, kind, and generous with your time and knowledge. I am extremely grateful to have been taught by such a great person and researcher as you.

To Constantine Sedikides, thank you for being my secondary supervisor and for letting me be part of the CRSI. You have gathered such an exceptional group of kind people, and I am so glad I got to know them and be a part of it. Thank you for being a great leader and always making everyone feel included. You are the funniest person I know!

To my parents, Jan Egil and Unn Levorsen, thank you for always supporting me in every way. I am very grateful to have you, and I would not have been able to do the PhD without you.

To the girls I met in the CRSI office - Carmen Surariu, Chloe Webb, Irem Ozdemir, Irem Yilmaz, Lily Huang, Monica Sood, Natalie Berry, Rhia Perks, and Wenjin Liu. I love you all so much and really appreciate that I got to do this with you guys - including all the coffees and gossip.

To Ayahito Ito, thank you for all the journal clubs and for being the best tour-guide in Japan. I really appreciate it.

To Andreea Butnaru, Ovidiu Brudan and Klara Dolaklova, thank you for being great friends and for all the coffees, phone calls and lunches. I am so grateful to have had you by my side throughout the PhD.

To Kailash Sharma, thank you for always looking after me, and for treating me as a part of your family during my time in Southampton.

Note on Formatting

The manuscript presented in Chapter 2 is published. The manuscript presented in Chapter 4 has been submitted for publication and received an invited resubmission. I have made two changes to these two original manuscripts to comply with the University's thesis guidelines. First, I changed the numbering of tables and figure. Second, I converted American spelling to British English.

Chapter 1 Literature review

1.1 What Is the Self?

Humans, as a social species, are dependent on social interactions for survival (Sedikides et al., 2006). To successfully navigate social interactions, they need knowledge about themselves (Decety & Sommerville, 2003). From an evolutionary perspective the self has evolved in response to human's ecological and social pressures (Sedkides & Skowronski, 1997). Sedikides and Skowronski (2000) propose an evolutionary account of the self, that is, the symbolic self. The symbolic self consists of three important capacities. The first capacity of the symbolic self is the representational self which stores an individual's mental representations of the past, present and future self. These representations include an individual's knowledge about themselves such as their traits, roles, values, preferences and beliefs (Markus, 1983). The second capacity of the symbolic self is the executive self and entails decision-making and the ability to regulate the self in relation to the physical and social environment. The executive self is guided by three motives: valuation (i.e., self-enhancement and self-protection), learning (i.e., seeking accurate information about the self) and homeostatic (i.e., seeking self-consistent information). The third capacity of the symbolic self is the reflexive self which involves an individual's ability to be conscious of themselves. This reflexive capacity allows an individual to depict themselves in the context of other objects and to flexibly change in response to the environment, such as altering long-term goals. Together, the dynamic interplay between these three capacities makes the symbolic self (Sedikides & Skowronski, 2000).

The current thesis focuses on the representational capacity of the symbolic self. I define the self-concept (or "the self") as knowledge about one's personality, such as our traits, roles, values, preferences, and beliefs (Markus, 1983). The self has been of interest to psychologist for over a century (James, 1890). For the past 25 years, the self has also been studied extensively by cognitive neuroscientists, who have consistently identified activation in the medial prefrontal cortex (mPFC) and posterior cingulate cortex (PCC) during self-related processing (Wagner et

al., 2012, 2019). Recurring questions have been whether the self is special in the brain, and what information about the self that is represented in the mPFC. Answers, however, remain elusive.

1.2 The Self as a Memory Structure

In the 1950s and 1960s, human cognition was primarily understood through two main approaches. One approach was based on a computer metaphor. According to it, cognition functions as a production system in which processes are activated through sequential, rule-based steps (Newell & Simon, 1972; O'Reilly & Munakata, 2000). Another approach proposed that cognition operates as a network of interconnected neurons (O'Reilly & Munakata, 2000). This proposal arose from advancements in neurobiology's understanding of neuron's functions. Examples include McCulloch and Pitt's (1943) binary model of neural processing, suggesting that neural firing is "all or nothing" and can be modelled as basic logical operations, and Hebb's (1949) finding that neurons that are coactivated develop stronger connections.

In the 70s, Anderson and his colleagues (Anderson, 1976; Anderson & Bower, 1973; see also: Anderson, 1983; Anderson & Conway, 1993) proposed the Adaptive Control of Thought model, which was based on both the computer metaphor and the neuron-like network framework. The model's aim was to explain human cognition in terms of language processing, inference making, and memory. The model included a production system that used if-then rules to perform procedural knowledge tasks, whereas an associative memory network served as a model for declarative knowledge. Procedural knowledge consists of "action-taking" productions that can be either mental or motoric skills (Winograd, 1975). Declarative knowledge consists of all factual information that people have, and can be divided into semantic and episodic memories.

The self-concept consists of knowledge about ourselves and is thus part of our broader social knowledge, which in turn forms part of the entire memory system. Specifically, the self-concept is a subset of our declarative memories of our social world (Kihlstrom & Cantor, 1984).

Here, semantic memory includes implicit personality theories, categorical knowledge about people and social situations, and detailed representations of both the self and other people in one's life. Whereas episodic memory comprises autobiographical memories of specific events and experiences (Kihlstrom & Cantor, 1984; Tulving, 1983).

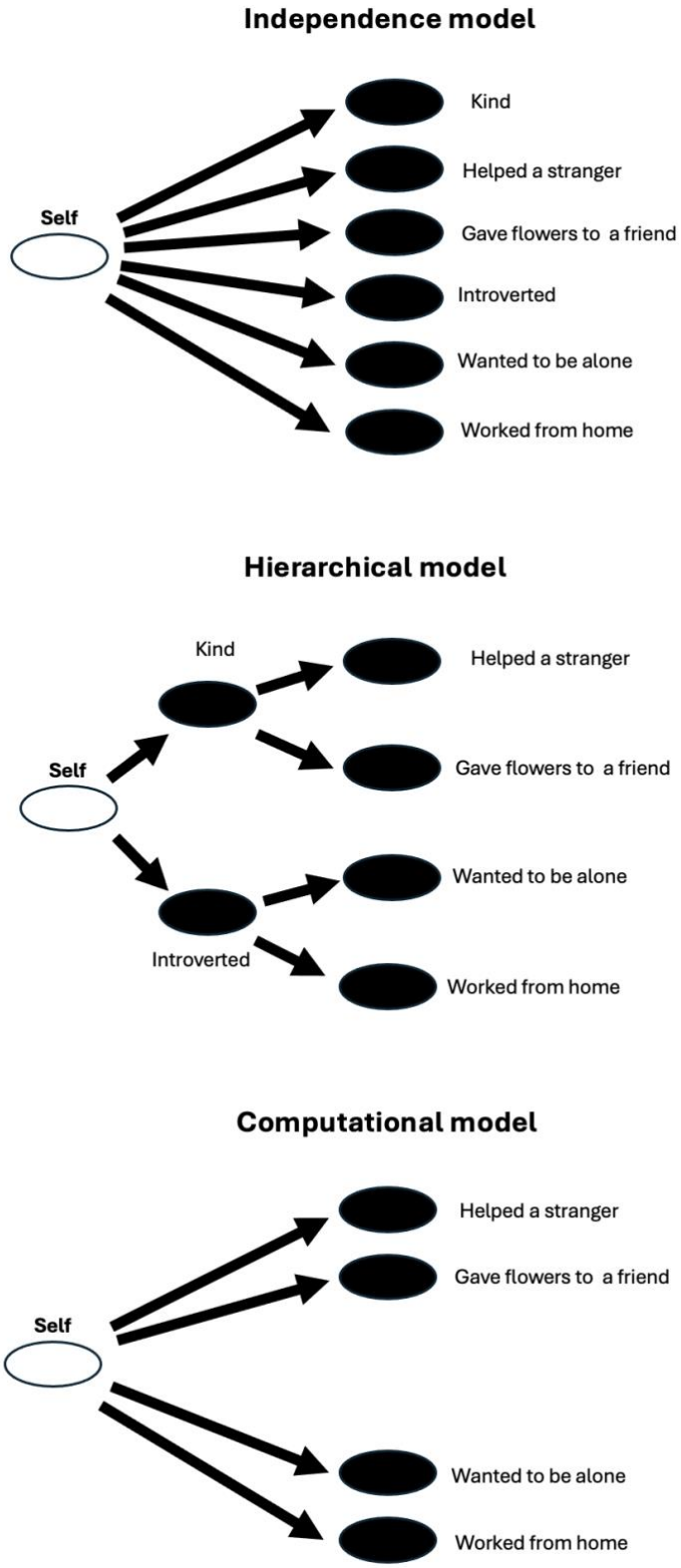
Given that the self-concept is part of memory, it can be represented as an associate network memory structure of one's declarative memories. Indeed, many psychologists have proposed associative network models of the self-concept (Bower & Gilligan, 1979; Greenwald et al., 2002; Kihlstrom & Cantor, 1984; Kihlstrom & Klein, 1994; Markus & Sentis, 1982). In an associative network model, the self is represented as a central node connected by associative links to other nodes that represent self-related concepts.

Three models have been developed to explain how associative network models are organised in memory (Kihlstrom et al., 2003). The independence model advocates that a self-node is connected to various separate nodes representing self-related concepts, both in terms of semantic knowledge and episodic knowledge (Anderson,1976; see: Sedikides & Green, 2000; Sedikides et al., 2016). The hierarchical model proposes a hierarchical organisation (Kihlstrom & Klein, 1994; Ostrom et al., 1980). Here, the self-node is directly connected to semantic knowledge nodes, which are themselves connected to various episodic nodes representing the specific semantic knowledge. Thus, the episodic knowledge nodes are only indirectly connected to the self-node. Lastly, the computational model posits that semantic self-related memories are not stored in memory, but are rather computed online based on episodic memories when needed (Anderson, 1974, 1981; Locksley & Lenaur, 1981). For example, when judging whether "kind" is self-descriptive, a person would first retrieve a memory of acting kindly and then complete the judgment based on the retrieval. (See Figure 1.1 for illustration of the three models).

In a competitive test of these models, Klein et al. (1989) found support for the independence model. Their research was based on the idea that a node in an associative link

can be activated by external stimuli and will spread activation to interconnected nodes (Anderson, 1976). The spreading of activation between nodes can be measured with reaction time latencies. Klein et al. presented participants (*N* = 24) with traits under a combination of three instructional sets: (1) to judge whether the trait described them (semantic memory); (2) to define the trait (control condition); (3) to recall a time they behaved in accordance with the trait (episodic memory). For each trait presented, participants carried out two of the three instructional sets; for example, when presented with the word "shy" they would first recall a memory, then judge if it described them. The two instructional sets were presented in sequence in varied order: (1) define then recall; (2) describe then recall; or (3) recall then describe. Additionally, participants repeated the same judgement twice consecutively, as a control condition. According to the hierarchical model, given that episodic memories are only indirectly linked to the self node through the semantic nodes, activating semantic nodes would facilitate judgements of episodic memory. Thus, making trait judgements would prime and facilitate subsequent performance in the recall condition. According to the computational model, recall of an episodic memory of a behaviour would prime and facilitate subsequent trait judgements. Finally, the absence of a priming effect would favour the independence model. Klein et al. obtained priming effects for a trait when the same task was repeated twice. However, they did not observe priming effects when the two consecutive trials consisted of different tasks, in support of the independence model. Follow-up research reinforced their findings (Klein & Lofthus, 1990, 1993; Klein et al., 1997).

Figure 1.1 Three Associative Memory Models

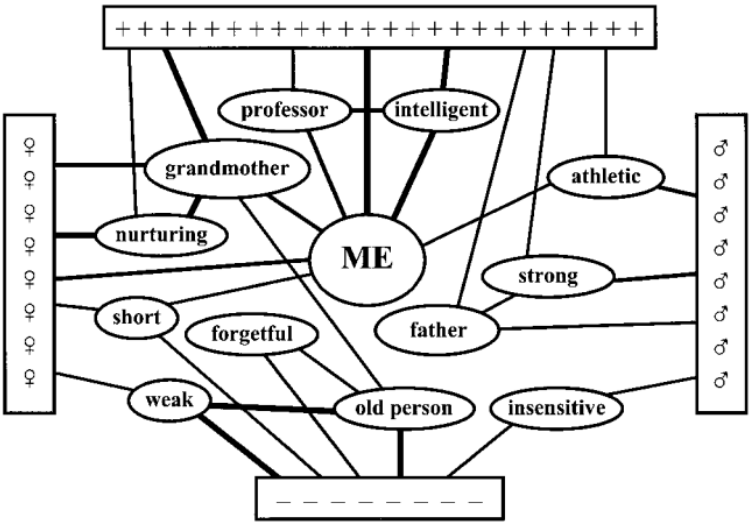


Note. Adapted from Kihlstrom and Klein (1994).

In the independence model tested by Klein et al. (1989), the self-related nodes were not interconnected. However, according to other authors, the nodes are also interconnected with one another (Greenwald et al., 2002; Kihlstrom & Cantor, 1984). For example, in a model proposed by Greenwald et al. (2002) (see Figure 1.2), links are interconnected to various degrees (represented by thickness), depending on strength of the association between concepts. If there is a strong associative link between two concept nodes, and one concept node is activated (either by external stimuli or another concept), activation will spread faster.

Strength of associative links have been measured with the Implicit association test (IAT; Greenwald et al., 1998, 2002). In the IAT, participants sort items into categories. The task is based on the assumption that, if words that are strongly related to each other in the associative network (thicker links) are sorted together, then reaction times will be faster. Research using this task has yielded numerous findings (see Fazio & Olson, 2003, for a review). Yet the psychological meaning of "thickness" of the associative links in the model remain unclear.

Figure 1.2 Associative Network Model of Self (Greenwald et al., 2002)



In summary, drawing inspiration from the functions of neurons, researchers developed associative network models to represent knowledge and the self-concept in memory. Researchers have also addressed whether the self, compared to other concepts in memory, has special properties. I turn to this issue next.

1.3 Self-Reference Effect in Memory

Given that a sense of self is essential for human experience, some authors have proposed that self-related stimuli are unique and have special properties in the brain, as they are especially memorable compared to other stimuli. Rogers et al. (1977) investigated whether there is a memory advantage for stimuli judged in relation to the self. They employed an extended version of the depth of processing task, previously used in memory experiments (Craik & Tulving, 1975) to examine the encoding strength of different conditions. In those memory experiments, the conditions included judging various properties of words, such as rhyme, structural properties, or meaning. The typical finding was that words judged for their meaning (semantic condition) were better remembered in a later recall task compared to the other conditions. Craik and Tulving suggested that the semantic judgement condition showed better recall, because it involved elaborative processes in memory; that is, when a word is judged for its meaning, it will be connected to related information in memory (creating traces), and these connections will later serve as retrieval routes. Words that create the most elaborate traces in memory are more deeply encoded and better remembered.

In Rogers et al.'s (1977) experiment, participants (N = 28) completed the self-reference task where they were instructed to judge if a word described them (e.g., Does "kind" describe you?), in addition to a semantic control (e.g., Does "happy" mean the same as "optimistic"?), phonetic (e.g., Does "shy" rhyme with "try"?), and structural (e.g., Is the word "caring" written in big letters?) condition. Following the judgement task, participants were asked to write down as many words as they could remember. They remembered more words judged in relation to the self compared to words judged in the semantic control condition (which had been found to show the highest recall; Craik & Tulving, 1975). As the self-reference condition resulted in better recall than even the semantic condition, and the main difference between the two conditions was self-reference processing, Rogers et al. concluded that the self has superior mnemonic abilities. Furthermore, they concluded that self-reference processing is powerful and rich for encoding information, leading to deep, elaborate traces in memory. They labelled the enhanced

recall of self-related stimuli as “the self-reference effect.” A limitation of this pioneering experiment was that it did not compare the self condition to another person. Thus, the effect could be due to judgements about social stimuli in general rather than the self in particular.

Bower and Gilligan (1979) compared the self-reference task to that of a close other. Here, participants (*N* = 40) viewed traits and completed four judgements: (1) “can you access a personal experience in which you exemplified this trait?”; (2) “Does the trait describe you?”; (3) “Can you access an incident, either directly experienced by you or told to you, in which your mother exemplified this trait?”; (4) “Does this adjective describe Walter Cronkite?”. (Walter Cronkite was a well-known journalist at the time). Next, participants engaged in a distractor task (addition calculation) for 10 minutes, followed by a recall task, where they were asked to list as many words as possible. Finally, they viewed all the traits again, in addition to new traits, in random order. Here, the task was to recognise the previously presented traits. Bower and Gilligan found enhanced recall and recognition in the self-reference condition, the personal memory condition, and the mother-reference condition, compared to the public other condition. There was no difference in recall or recognition, across the self-reference, personal memory, or mother-reference conditions. The results, then, indicate a reference effect for both the self and a close other, contradicting Rogers et al.’s (1977) conclusion that the self is a unique cognitive structure. According to Bower and Gilligan, both the self-reference and other-reference effects are due to linking presented traits to a pre-existing knowledge structure. This linking is done more efficiently in these two tasks, compared to the semantic condition, because people have relatively elaborate knowledge about the self and close others. Following Bower and Gillian, a large literature replicated the self-reference effect both in comparison to a semantic control condition and another person (Symons & Johnson, 1997).

Klein and Kihlstrom (1986) argued that the self-reference effect is a result of differential organisation of stimuli that occurs during the self-reference task compared to control conditions. Their argument is based on the notion that judgements made in terms of categories are organised differently in memory compared to words that are not judged in relation to

categories. In the self-reference task where participants judge if a word describes them, the items are divided into two categories: words that are self-descriptive and words that are not. However, this is not the case for judgements made in the control conditions (e.g., "Is the word written in capital letters?"). Klein and Kihlstrom proposed that this difference explained the enhanced recall for self-reference judgements. They conducted an experiment to test this hypothesis. Participants (*N* = 64) were randomly assigned into four conditions: semantic-unorganised, self-unorganised, semantic-organised, self-organised. Participants in the semantic-unorganised condition read incomplete sentences, followed by a target word, and judged if it was sensible to use the target word to fill in the gaps in the sentence (e.g., "The soldier preferred to keep his ___ short, Hair?"). Participants in the self-unorganised condition also read incomplete sentences and a target word, and judged if the sentence described them (e.g., "I always keep a civil__, Tongue?"). In the semantic-organised condition, participants judged if a target word was an external body part ("Is this an external body part, Heart?"). In the self-semantic condition, participants judged if they had ever had an injury related to the presented word (e.g., "Can you think of an incident in which you had an injury or illness associated with your Leg?"). Here, the two questions in the organised condition could be answered by dividing the response into two categories (internal/external body parts and self-descriptive/not self-descriptive), whereas this was not the case for the unorganised conditions. Klein and Kihlstrom maintained that words in both organised conditions were part of a sub-group (based on the categories), whereas words in the unorganised condition were not part of a sub-group. After the judgement task, all participants wrote down as many words as they could recall. The results indicated enhanced recall in both organised conditions compared to the two unorganised conditions. There was no difference in recall between the two unorganised conditions. Hence, the researchers concluded that the self-reference effect is a result of organisation rather than the self being a superordinate, highly elaborate construct in memory.

Klein and Lofthus (1988) reported that both elaboration and organisation play a role in self-reference processing. In prior work, when words in a list were inter-related, an elaboration

task, compared to an organisation task, lead to better recall, given that the stimuli were organised to begin with; also, when words were unrelated, organisation, rather than elaboration, produced higher recall (Begg, 1978; Klein et al., 1988). Klein and Lofthus reasoned that, if they compared both an organisation task and an elaboration task to the self-reference task, they could examine the relative contribution of elaboration and organisation. The experiment involved a 3 (organisation task, elaboration task, self-reference task) x 2 (related words list, unrelated words list) between-subjects design. In each task, participants (N = 84) were first presented with 30 cards that were placed upside down. Each card contained a word. Participants picked up the cards one at a time and read the words sequentially. In the elaboration task, they generated a definition of the word and rated how difficult it was to do so. In the organisation task, they placed the word into one of five given categories. In the self-reference task, they thought of an experience they had that was related to the word. A distractor task and recall followed. The results indicated that, when stimuli were related, the self-reference task functioned as an elaboration task, that is, recall was similar in both the elaboration and self-reference conditions. Recall was better in the elaboration task and the self-reference task than in the organisation task. On the other hand, when stimuli were unrelated (and organisation was needed to enhance recall), recall did not differ between the self-reference task and organisation task. Recall was better in the self-reference task and organisation task than in the elaboration task. In interpreting the results, Klein and Lofthus advocated a dual-process account, suggesting that elaboration and organisation are separate processes implicated in self-reference. However, which of the two processes will have the largest impact on recall depends on task material (i.e., extent to which stimuli can be related to each other).

In an attempt to integrate the literature on the self-reference effect, Symons and Johnson (1997) conducted a meta-analysis. They included 42 experiments that compared the self-reference task with an other-reference task (involving either a non-close or close other) and a semantic control tasks. They reported a robust self-reference effect, both compared to other-

reference and semantic control conditions. When they compared the effect of self-reference to an other-reference with a close other, the effect was smaller compared to the effect of both a non-close other and a semantic control condition. Consistent with Klein and Lofthus's (1988) conclusion above, they explained their findings in terms of elaboration and organisation. They argued that self-reference and close-other reference processing have an inherent organisation advantage over semantic control conditions, because they can be judged in relation to categories (e.g., self-descriptive and non-self-descriptive words). Symons and Johnson also highlighted that the robust self-reference effect is a consequence of elaboration: when judging the meaning of a word, one relates it to already existing semantic memory that is rich in the case of self-knowledge. Additionally, these authors emphasised that people think about themselves frequently, thus self-reference is more frequently practiced compared to control conditions. Symons and Johnson concluded that self-reference is unique in the way that it is especially often thought about, more elaborated and can be organised in terms of two categories in memory compared to any other information. They also highlighted that it would be premature to conclude that the self is a unique cognitive structure in memory.

Taken together, early research on the self-reference effect has converged in showing enhanced recall for self-related judgements compared to semantic judgement. To explain this effect, three explanations have been proposed: (1) the self is a unique cognitive structure in memory; (2) people have elaborate self-knowledge, just as they can have elaborate, close other-knowledge; and (3) the self-reference effect arises from a judgement that can be categorically organised. The results of behavioural experiments, then, have been inconclusive. I turn to the neuroimaging literature.

1.4 The Self-Reference Memory Effect in Neuroimaging

Armed with neuroimaging methods, cognitive neuroscientists ventured into the investigation of the self-reference memory effect. In the late 90s, a group of researchers at the University of Toronto in Canada, conducted the first experiment (N = 8) of the neural basis of the

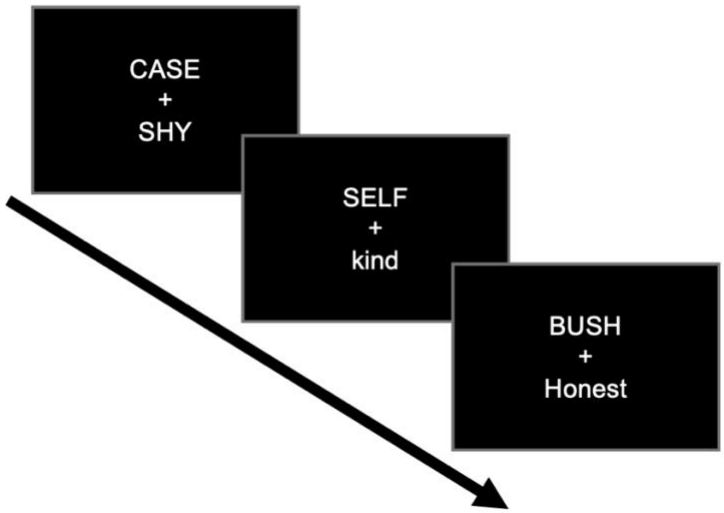
self (Craik et al., 1999). They used positron emission tomography (PET) to scan participants while they performed a judgement task. Participants viewed trait adjectives and judged them in four conditions: (1) self-reference task ("How well does the adjective describe you?"); (2) other-reference task ("How well does the adjective describe Brian Mulroney?" – the former prime minister of Canada); (3) social desirability ("How socially desirable is the trait described by the adjective?"); (4) and number of syllables ("How many syllables does the adjective contain?"). The first three conditions required semantic judgement, whereas the fourth condition (baseline) required phonological judgement. After the scan, participants completed a surprise memory recognition test. They viewed all the words again and indicated whether they remembered them. The results did not reveal the standard self-reference effect. The experiment, though, has two limitations. First, the sample consisted of only eight participants. Secondly, as PET experiments require long intervals between trials, each participant completed a relatively low number of trials (32). These limitations may have led to low statistical power, preventing the experiment from detecting the self-reference memory effect.

Craik et al. (1999) conducted two neuroimaging analyses: one regular contrast subtraction analysis and a partial least square analysis. The latter is more sensitive, as it includes information about the covariance between voxels in the analysis. When comparing the neural activation across the conditions, the authors highlighted that the regular analysis indicated that activation was very similar for the three semantic conditions. However, when contrasting the self-reference condition to all other conditions in the partial least square analysis, they found activation in mPFC, suggesting that the self is distinctively represented in the mPFC.

Kelley et al. (2002) carried out an fMRI experiment with an event-related design and a larger sample size (N = 22) of US participants. While in the scanner, participants completed the self-reference task, the other-reference task, and a case judgement (control condition). In the other-reference condition, participants judged if a trait adjective described George Bush (the president at the time), whereas, in the case control condition, they judged if a word was written

in upper-case letters (Figure 1.3). After the scan, participants completed a surprise memory recognition test. Kelley et al. replicated the self-reference effect: words judged in relation to the self were better remembered compared to the two other conditions. Importantly, mPFC and PCC showed greater activation in the self-reference condition compared to the other two conditions. Although the authors did not test if the greater mPFC activation on self-reference trials was related to the enhanced behavioural memory effect, they interpreted their results as preliminary evidence of an involvement of the mPFC in self-referential processing. They concluded that self-reference processing is dissociable from other processes in the brain.

Figure 1.3 Example of the Self-Reference and Other-Reference Task



Note. Adapted from Kelley et al. (2002).

Macrae et al. (2004) carried out an fMRI experiment in which they found support for the idea that mPFC activation is responsible for the self-reference effect. During the scan, participants (N = 22) made self-reference judgements ("Does this adjective describe you?") in addition to control trials where they fixated on a central point on the screen. After the scan, participants went through a surprise memory recognition test. When comparing activation during the self-reference task to the control trials, Macrae et al. found an increase in activation in the visual cortex, parahippocampal gyrus, parietal cortex, dorsal prefrontal cortex, and the cerebellum. They also found a decrease in activation in the PCC, mPFC, and right

hippocampus. When they compared activation during trials with words that were later remembered to trials of words that were later forgotten, they found greater activation in the mPFC, anterior prefrontal cortex, and parahippocampal gyrus. As mPFC was recruited during self-reference judgements of remembered trial, Macrae et al. concluded that self-processing is dissociable from other processes and thus supported the idea that the self is "special" in the brain. However, this conclusion might have been prematurely drawn. Because the authors only compared the remembered trials with the forgotten trials for the self-reference condition, without comparing the effect to another person, they could not conclude that their results were specific to the self. As reviewed above, Bower and Gilligan found enhanced memory recall and recognition in the close other condition relative to the semantic condition; thus, it would be informative to examine whether this is also the case when relating the self-reference effect to neural activation.

After Macrae et al.'s (2004) inital finding that the mPFC is responsible for the self-reference effect in memory, only a few studies have aimed to replicate it. Kim and Johnson (2012) used fMRI to measure brain activation while participants performed an object allocation task. Specifically, participants (N =12) were presented with two baskets, one on the left side and the other on the right side of the screen. One basket was named "Mine" and was a certain colour, whereas the other basket was named "Alex" (stranger) and was another colour. For each trial, a new object was presented, and a coloured dot appeared to indicate the basket in which the object should be assigned. Participants pressed a left/right button depending on the colour of the dot. After the scan, participants underwent a memory test: they were shown each object again and indicated if it had previously been assigned to the "Mine or "Alex" basket. The behavioural results of the main task showed that objects assigned to the self were remembered better than objects assigned to the other. Further, the results indicated greater activation in the mPFC, paracingulate gyrus, and frontal pole for objects assigned to the self than object assigned to the other. Furthermore, activation in the mPFC was greater for remembered words in the self condition compared to the other condition. Kim and Johnson did not use the same

contrast as Macrae et al. that had compared neural activation for remembered words to forgotten words. However, the Kim and Johnson results illustrated that the mPFC is related to a self-reference effect also in terms of object ownership. Although this experiment included an other-person condition, it remains unknown how mPFC activation is related to the self-reference memory effect when compared to a close other.

On the other hand, Koski et al. (2020) failed to replicate the findings by Macrae et al. (2004) and Kim and Johnson (2012). They followed the task design and analyses contrasts of Macrae et al. (2004). However, when comparing remembered to forgotten words, they did not find any significant activation in the mPFC.

Relatedly, a lesion study investigated if the mPFC is necessary for the self-reference effect and found evidence suggesting that it is. In particular, Phillippi et al. (2012) compared performance on the self-reference task for six patients with mPFC focal damage to performance of patients with lesions to other cortical (i.e., lateral occipital, temporal, and parietal) regions, as well as healthy control participants. Patients with mPFC lesions showed a weaker self-reference effect compared to both the brain damage group and healthy controls, suggesting that the mPFC is necessary for a consistent and reliable self-reference effect.

The above reviewed experiments have limitations. Although the Philippi et al. (2012) findings suggest that the mPFC is necessary for a consistent and reliable self-reference effect, the null findings of Koski et al. (2020), combined with the low sample size (N =12) in the Kim and Johnson experiment (2012), raise doubts as to whether activation in the mPFC predicts the self-reference memory effect. Furthermore, it is unclear whether the self-reference effect is specific to the self, due to lack of a control condition involving a close other.

In summary, neuroscientists used PET and fMRI to investigate the self-reference effect. Compared to control conditions, the mPFC showed greater activation during the self-reference condition. A few experiments reported a correlation between the mPFC activation during the self-reference task and the behavioural self-reference effect. Also, a lesion study found that

mPFC is necessary for the self-reference effect. However, these experiments were few and have limitations. A persistent question in the neuroimaging literature on the self revolves around understanding if the self is special in the mPFC. Consequently, a large amount of research has compared activation during the self-reference task to that of the other-reference task. I review this research next.

1.5 Comparing the Neural Basis of Self and Other

1.5.1 Self-Reference versus Other-Reference

Social neuroscientists have focused not on the causes of the self-reference memory effect, but rather on how the self is represented in the brain. Thus, numerous neuroimaging experiments have used the self-reference task, however, they have not incorporated a surprise memory test, because their objective was simply to identify brain regions activated during the self-reference task.

Most of such neuroimaging experiments have compared the self-reference task to an other-reference task, where the other varies in relation to the self on dimensions such as familiarity, similarity, or closeness. For example, in the Kelley et al. (2002) experiment outlined above, the self is compared to a famous person (George Bush), whereas in other experiments the self is compared to a similar stranger, similar friend, close friend, or family member (Krienen et al., 2010; Schmitz et al., 2004; Vanderwal et al., 2008). Experiments comparing self-reference judgement to judgement of other often use a famous public figure that participants know of but are not personally familiar with. The results usually demonstrate greater activation in the mPFC in the self-reference condition compared to the other condition. On the other hand, experiments comparing self-reference to similar or close others have produced mixed findings (Wagner et al., 2012).

One stream of literature suggests that similarity between self and others is related to activation in the mPFC during the other-reference task. In an fMRI experiment, Benoit et al.

(2010) asked participants (*N* = 13) to perform the self-reference task as well as the other reference task where they were asked to make descriptiveness judgements about their best friend. Both tasks involved four response options for the descriptiveness judgements: “yes sure”; “yes unsure”; “no sure”; “no unsure.” The correlation between responses in the self-reference task and other-reference task was used as the similarity measure. They did not find any difference in mPFC activation between the self and other conditions. However, they found an interaction between the self-other contrast and the similarity measure so that the higher the similarity between self and best friend, the stronger the activation in mPFC during the other-reference task.

In an fMRI experiment, Mitchell et al. (2005) also examined neural activation for similarity to the self. Participants (*N* = 18) viewed photographs of faces and made two types of judgements while in a scanner: a mentalising and a non-mentalising one. In the mentalising condition, they judged how pleased the person in the photograph was, whereas, in the non-mentalising condition, they judged how symmetrical the person’s face was. After the scan, they again viewed the faces and rated them on similarity to the self. The dimension of similarity to the self was not specified. Parametric modulation analysis indicated that the higher the similarity ratings, the greater the activation in the vmPFC during the mentalising task. Mitchell et al. (2006) conducted a similar experiment. Here, the similarity dimension was specified; they investigated similarity in terms of political orientation. Mentalising about politically similar others recruited the vmPFC, whereas mentalising about political dissimilar others recruited the dmPFC.

Further, Jenkins et al. (2008) used fMRI repetition suppression to investigate if a judgement of a similar and dissimilar other engages activation like a judgement about the self. Repetition suppression is based on the assumption that, when conceptually similar stimuli are presented several times in a row, neural activation responding to the stimuli is suppressed for the repeated stimuli (Grill-Spector et al., 2006). Firing of neurons for a stimulus is thought to be fatigued and thus reduced with repetition. Repetition suppression, then, can address if two stimuli activate the same or different populations of neurons in the brain. Participants (*N* = 13)

first viewed images and information about two people: one with similar and one with dissimilar political views to their own. Next, in an fMRI scanner, they judged how likely the similar other, the dissimilar other, and themselves would be to agree with a range of statements (e.g., "enjoy helping friends with problems"). Each trial was followed by either an identical trial or a trial for a different condition. In this way, researchers could test if the neural response was suppressed for self-self, similar other-self, and dissimilar other-self. A repetition suppression effect emerged in the vmPFC for self-self or self-similar other trials, but not self-dissimilar other trials. The results suggested that people rely on their own minds to understand those of similar others, but not dissimilar others.

To explain these findings, Mitchell et al. (2005) proposed a self-simulation process in the vmPFC. Simulation theory suggests that people use their own thoughts, feelings, and behaviours to understand others' mental states (Gallagher & Frith, 2003). People, then, use themselves as a model to understand others. However, this is only plausible if the other person is similar to oneself. Given that the vmPFC is active during both self-reference processing and while thinking of similar others, the abovementioned experiments suggested that activation in the vmPFC during judgements of similar others reflects self-reference processing taking place to infer and understand others' thoughts (Benoit, 2010; Mitchell et al., 2005, 2006). Benoit et al. (2010) further proposed that similarity, and hence the degree of simulation, could explain the mixed findings of the mPFC activation overlap when comparing self-reference and other reference. Thus, it may be the variation in similarity to the self that is reflected in the mPFC activation during the other-reference task.

Inconsistent with the idea of a self-simulation process in the mPFC (Mitchell et al., 2005), Krienen et al. (2010) reported that close others, but not similar others, showed an overlap in mPFC activation with self-reference. Before an fMRI experiment, participants (N = 28) provided photographs and information about two of their friends, one who was similar to them, and one who was dissimilar. Participants were also presented with photographs and information about a similar and dissimilar stranger. Similarity to the self depended on individual biographic

information provided by each participant prior to the scan. It included details about personality, education, and lifestyle. In an fMRI scanner, participants viewed photographs of either themselves, the similar friend, dissimilar friend, similar stranger, or dissimilar stranger. They were presented with a statement of a personal preference and asked to judge what they believed the preference for the person in the photograph would be. Krienen et al. found greater activation in the mPFC for self and close others (both similar and dissimilar) compared to strangers. However, there was no difference in activation between self and close others. These results indicate an overlap in activation for self and close others, regardless of similarity. The authors proposed that, instead of similarity to the self, mPFC activation reflects the significance or social value to the self. That is, due to its importance for survival, people have evolved mechanisms to discern social information that is relevant to them, such as identifying information about close others.

In that vein, an experiment by Heleven and Van Overwalle (2019) suggested that the mPFC responds to both self and close others. Similarly to Jenkins et al. (2008) mentioned above, Heleven and Van Overwalle used fMRI repetition suppression to compare neural responses while participants thought about self and a close other. Participants (N = 28) performed the self-reference task and other-reference task. When completing the other-reference task, they thought about people close to them (e.g., friend, family member). The conditions were: self followed by self; self followed by a close other; close other followed by self; a close other followed by another close other. Like Jenkin et al., Heleven and Van Overwalle observed a repetition suppression effect in the vmPFC when self-reference was preceded by other-reference. However, they did not observe the effect the other way around. The researchers reasoned that close others are likely to be similar to the self and so the asymmetrical results can be explained by feature of similarity theory (Tversky, 1977). This theory states that, when two similar concepts are compared, the one that is smaller (or people have less knowledge about) will seem more similar to the bigger concept than the bigger concept will seem similar to the smaller one. This is because people can intuitively picture the smaller concept to be a part

of the bigger, but not the reverse. In the context of Heleven and Van Overwalle's experiment, the close other might seem more similar to the self than the self seemed similar to the close other.

Results from additional experiments indicate that the mPFC shows no difference in activation when thinking about self and a close other, whether in terms of personality traits or food preferences. For example, Schmitz et al. (2004) and Ochsner et al. (2005) compared self-reference judgements of trait adjectives to those of a close friend. They found no difference in mPFC activation between the two conditions. Similarly, Vanderwal et al. (2008) reported no difference in mPFC activation when participants compared trait judgments of themselves and their mother. Seger et al. (2004) compared judgements about food preference for self and a close other, and observed no difference in mPFC activation between the two conditions. Furthermore, Moran et al. (2011; $N = 21$) compared self-reference judgements to other-reference judgements of one's mother in regard to traits and physical appearance. They found no difference in mPFC activation for self and mother in terms of traits, but greater activation in the mPFC for appearance judgements in the self condition. In a follow-up experiment, they asked participants to rate the same aspects on importance to the self. Consistent with their neuroimaging results, participants' own traits and appearance, in addition to their mother's traits, were most important to the self. Moran et al. concluded that mPFC activation reflects the integration of a close other's personality into one's self-concept (Aron et al., 2004).

However, other experiments have reported contradictory findings, namely, that mPFC activation is greater for thinking of the self than a close other. Heatherton et al. (2006) asked participants ($N = 30$) to judge if traits described themselves, a close other (i.e., their best friend), or was written in upper case letters. They observed greater mPFC activation during self-reference judgments compared to both the close other and a control (i.e., case judgement) condition. D'Argembeau et al. (2007, 2008) found similar results. Also, Chen et al. (2013) reported greater mPFC activation during self-reference judgements compared to other-reference judgments about one's mother. In all, despite extensive research (Wagner et al.,

2012), experiments that contrasted neural activation during the self-reference task to activation during the other-reference task with a close other have yielded mixed results.

In a meta-analysis of 25 experiments, Murray et al. (2012) compared experiments that used the self-reference task to experiments that used the other reference task in which the other person was either a close other or a famous person. When contrasting self to both close other and a familiar other, Murray et al. found activation in a cluster in the mPFC, suggesting that an area of the mPFC is self-specific. Further, when they compared self, close other, and public figures separately to a semantic control condition, they found a substantial activation overlap for self and close other in the vmPFC. They additionally observed a partial overlap in activation for self and famous persons in the mPFC: whereas activation for self and close other was located in the ventral part of the mPFC, activation for a famous person was located more dorsally. The authors proposed a reinforcement-expectancy account to explain vmPFC activation. According to this account, based on previous experience, the vmPFC processes the anticipation of rewards and exhibits sensitivity and bias toward self-related aspects of stimuli. Given that the vmPFC was active while thinking about the self and close other, but not a famous person, this region might track inherent self-relevant features that evoke reinforcement expectancy responses.

In another meta-analysis, Qin and Northoff (2011) included experiments from four conditions: self; close other (friends or family); famous person or stranger; resting state. They included data from 23 experiments for each condition. In contrast to Murray et al. (2012), who only included experiments related to trait judgements about the self, Qin and Northoff included experiments using a wider range of self-related stimuli, such as self-face recognition, agency, sensory experiences, body experiences, thinking of the past and future, as well as personality traits. Qin and Northoff located a cluster in the mPFC that was active specifically for the self when contrasted with the close other conditions. Additionally, they located a small cluster with overlapped activation for the self and close other in the mPFC and PCC. Activation for the self and resting state also overlapped in the mPFC. Taken together, these two meta-analyses

including different types of self-related stimuli, both found areas within the mPFC that are self-specific, whereas other areas within the mPFC are also active when thinking of close others.

Confounding variables may explain the mixed findings of mPFC activation for self versus similar or close other (Gillihan & Farah, 2005; Koski et al., 2020). For example, the person in the other condition was often someone very close to the self (e.g., mother, partner, relative, best friend). Thus, it is likely that participants had a lot of memories, previous knowledge, and personal familiarity for a close other compared to a stranger (Koski et al., 2020). Although it is possible that the factors proposed above (closeness, similarity, or value) explain the mPFC activation when comparing the self and other reference task, the extent to which confounding variables are driving these results is unknown.

Another explanation for the mixed results in experiments comparing self-reference to other-reference is that both tasks require general cognitive processes that recruit mPFC but that are unrelated to the self (Legrand & Ruby, 2009). Building on the early findings of a self-reference memory effect (Rogers et al., 1977), authors of several neuroimaging experiments (Kelley et al., 2002; Macrae et al., 2004; Ochsner et al., 2005) concluded that the mPFC and PCC are a self-specific neural network. Legrand and Ruby (2009) disagreed with this conclusion. They highlighted that a range of other tasks (i.e., theory of mind, memory recall, resting state, inductive and deductive reasoning) also activate the mPFC, and proposed that this activation might be a result of general cognitive processes that commonly take place during various tasks. Other reviews concurred (Gillian & Farah, 2005; Ruby & Legrand, 2007). Also, experiments demonstrated that processes such as autobiographical memory (Martinelli et al., 2013) and reward (Bartra et al., 2013) activate similar regions within the mPFC. In other words, the bulk of the literature addressing the self in the brain focused on the link between mPFC activation and the self-reference task rather than the link between mPFC activation and cognitive processes. As such, they did not control for potential confounds involved in the self-reference task. It remains unknown, then, whether the self is indeed “special” in the brain.

In contrast to fMRI experiments, lesion studies can provide information regarding the necessity of mPFC for self-reference processing (Vaidya et al., 2019). Marquine et al. (2016) showed that the mPFC is critical for retrieval of personal trait knowledge of self, but not of a close other. They conducted a case study of patient J.S., who sustained a bilateral lesion to the mPFC. J.S. rated the extent to which traits (e.g., moody, considerate) applied to himself and another person (i.e., a nurse who had worked with him for nine years). The researchers compared J.S's ratings to those of healthy control participants' ratings of themselves and a close other. All ratings were carried out on two separate days, 10 days apart. J.S.'s close other (i.e., the nurse) also rated the extent to which the trait words described J.S. and themselves. Healthy participants' close others did the same. Accuracy scores were calculated as the intraclass correlation between J.S.'s self-ratings and those of the nurse. Consistency scores were calculated as the intraclass correlation between J.S's first and second rating of himself. Both in terms of accuracy and consistency, J.S.'s trait-knowledge was severely impaired compared to controls. However, both in terms of accuracy and consistency, his trait knowledge of the close other (i.e., nurse) was intact. Although this study only included one patient, the results indicate that the mPFC is necessary for processing knowledge of the self, but not of others, suggesting a specialised role of the mPFC in self-reference processing.

Taken together, the literature comparing mPFC activation for thinking of the self and a close other have produced mixed results. Explanations of similarity, closeness, and social value to the self have been proposed to explain the inconsistent results. However, it is also likely that the results are driven by confounds such as previous knowledge and personal familiarity. Moreover, the mPFC activation during both the self and other-reference tasks might be due to common general cognitive processes taking place during the task. Despite a number of suggestions, and extensive research, the explanations of the mixed findings for self-reference versus other-reference remain insufficient. Although the fMRI experiments could not provide any answers regarding specificity, a lesion study suggests a specialised role of mPFC in self-reference processing. Since these early studies were conducted, a more sensitive analysis

approach which can account for some of the abovementioned confounds has been developed. The following section will outline these analyses methods and studies utilising them to examine the representation of self and other in the brain.

1.5.2 Multivariate Analysis Experiments Comparing Self and Other

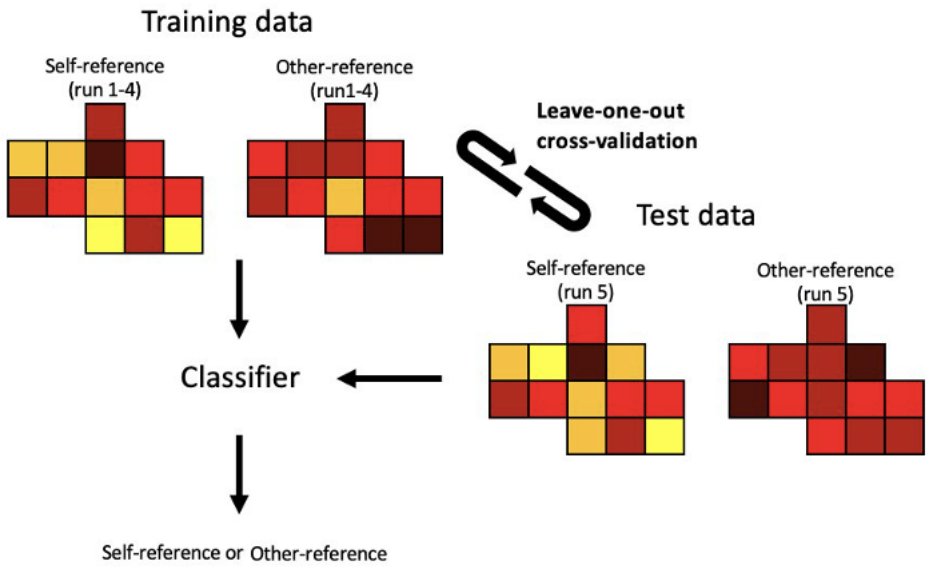
The experiments reviewed so far have used a univariate analysis approach. In the last decade, the neuroimaging literature on the self has increasingly shifted from a conventional univariate analysis approach to a multivariate analysis one (Wagner et al., 2019). Univariate analysis is run for each voxel in isolation, simultaneously. Then, the magnitude of activation in a region, or the whole brain, is compared for different conditions (e.g., self-reference versus other-reference; Norman et al., 2006). Although the research examining the self in the brain with a univariate analysis approach has been extensive (Wagner et al., 2012) and has identified regions that are active during the self-reference task, there is only so much information such an approach can provide. Multivariate analysis is more sensitive, as it includes information about the distributed patterns of activation across multiple voxels. As such, it moves beyond merely comparing the magnitude of activation for two condition, allowing the test of more complex questions (Dimsdale-Zucker & Ranganath, 2019).

Furthermore, in contrast to univariate analyses, multivariate analyses can provide information regarding a common neural mechanism of different tasks (Woo et al., 2014). For example, experiments using a traditional univariate analysis approach found that both social rejection and physical pain activate the dorsal anterior cingulate cortex (dACC; Eisenberger et al., 2003; Kross et al., 2011), implying that the two forms of pain are based on a common neural representation (Kross et al., 2011). However, multivariate analyses later revealed that these two tasks rely on separate neural mechanisms in the dACC (Woo et al., 2014). The same holds for social conformity and reinforcement learning: a shared neural mechanism for the two tasks was proposed based on univariate results (Klucharev et al., 2009), whereas distinct neural mechanisms were uncovered using multivariate analyses later (Levorsen et al., 2021). Put

otherwise, researchers cannot draw conclusions about a common neural representation based on a univariate analysis activation overlap. Next, I consider various multivariate analysis approaches.

Multi-voxel pattern analysis (MVPA) can include several different types of multivariate analyses; however, it is most often used with a multivariate pattern classifier. The runs of data are divided into training and testing subsets. For example, if the purpose is to investigate whether the self-reference and other-reference task are based on distinguishable patterns of neural activation, and the researcher has five runs of data, they would first assign the first four runs for both tasks as training data. The remaining data, run 5, would be used as test data. The classifier would first use the training data to attempt to learn to distinguish the data from the two tasks from each other. After training, the classifier would be applied to the test data to find out if it can reliably distinguish neural activation during the self-reference task from that of the other-reference task. If it can do so with an above chance accuracy, it is assumed that the two conditions are based on different neural mechanisms (Norman et al., 2006; Wagner et al., 2019) (See Figure 1.4). This type of analysis is often employed with leave-one-out cross-validation. That is, parts of the data are initially designated as training data, whereas other parts serve as test data. After running the analysis and calculating accuracy, the roles are systematically swapped until each run has been used as test data. For example, in the first round, run 1-4 is used as training data, whereas run 5 is used as test data. Subsequently, in the next round, run 2-5 would be used as training data, and run 1 as test data. In the following round, runs 3, 4, 5, and 1 would be used as training data, whereas run 2 would be used a test data, and so on. In this way, the test data are always independent from the training data.

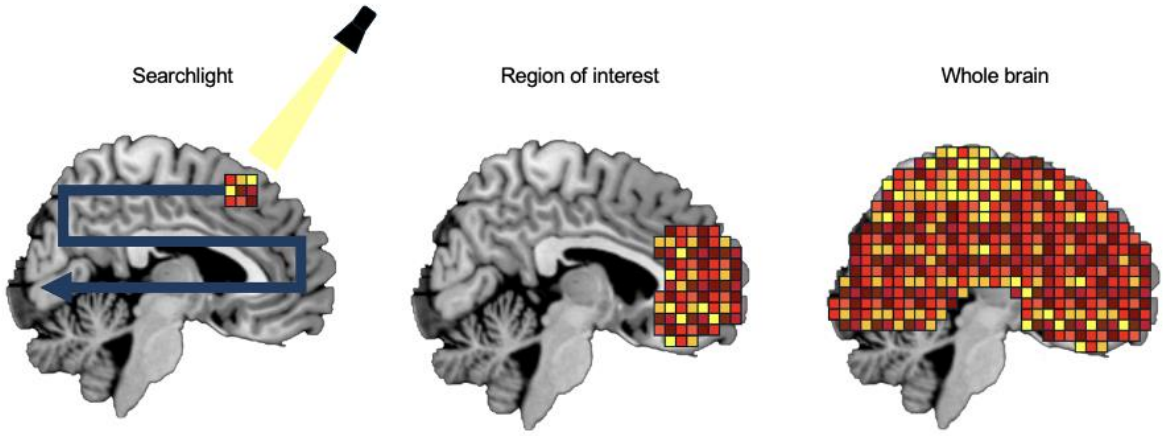
Figure 1.4 Multi-Voxel Pattern Analysis



Note. Each square represents the activation value of a voxel (the darker the colour, the higher activation) for the given condition. MVPA is conducted separately for each participant.

MVPA approaches can be applied at various spatial scales, such as searchlights, region of interest (ROI), and the whole brain (Figure 1.5). MVPA approaches are often used in combination with searchlight analysis (Kriegeskorte et al., 2006) to investigate the local patterns of activation. In searchlight analysis, a "searchlight" is put on a small subset of voxel at a time. Each voxel within the whole brain or an ROI is subsequently a centre voxel. For each searchlight the value from the centre voxel and its surrounding voxels are extracted, and the analysis (e.g., MVPA with a classifier) is conducted within the given searchlight. The searchlight is then moved to the next centre voxel where the same process is repeated. This process continues until all voxels in the brain or the given ROI have been a centre voxel. The shape and size of each searchlight varies across experiments. The shape is three dimensional and can be formed as a cube, a diamond, or a sphere (Dimsdale-Zucker & Ranganath, 2019). Searchlight analysis is the smallest spatial scale and can identify locally distributed patterns of activation. Searchlight analysis is useful to investigate psychological processes with fine-grain representations that might not be captured at the regional or whole brain level (Jolly & Chang, 2021).

Figure 1.5 Different Spatial Scales Used to Perform MVPA: Searchlight, Region of Interest and Whole Brain



Koski et al. (2020) used MVPA to compare the neural code for self-reference to other-reference in the mPFC. They used a minimal group paradigm to compare the neural representation of the self to a minimally related other (a person allocated to the same group as the participant prior to the scan) and an unrelated other. They intended to test if experiments that reported an overlap in mPFC activation for the self and both close and similar others were confounded by previous knowledge. They therefore included a minimally related other condition, a condition that is emotionally related to the self, but with no previous knowledge. Prior to an fMRI scan, participants (N = 48) viewed pictures of two strangers who ostensibly had taken part in this experiment and were randomly assigned to a red or a blue group. Next, participants were led to believe that they were randomly assigned to the blue group, and thus shared a minimal connection with the person in their group. Participants performed both the self-reference task and the other-reference task in which they made judgements about the two others: the minimally related other and the unrelated other (the person who was assigned to a different group). MVPA with a searchlight approach was performed. The searchlight was moved throughout a functionally defined mPFC ROI. The MVPA was carried out for each possible pair of comparisons for the three conditions (i.e., minimally related other versus unrelated other, self versus unrelated other, self versus minimally related other). The results revealed different

patterns of activation in the vmPFC for all three conditions (self, minimally related other, unrelated other), suggesting that the mPFC distinguishes people based on specific identities.

Similarly, Parelman et al. (2022) used MVPA to compare the neural representation of self-reference to other-reference processing. They employed a mega-analysis approach combining the data of three fMRI experiments (*N* = 142). The first experiment examined the effect of self-affirmations on neural responses while reading threatening health message (Falk et al., 2015). The second experiment examined the relationship between neural activation during anti-smoking messages and subsequent behavioural change (Cooper et al., 2015). The third experiment (unpublished) involved measuring valence in addition to thinking about the self and other. Each experiment included both self-reference and other-reference tasks. The other person in the other-reference task varied across experiments: in two of the experiments the person was a famous other (i.e., Barack Obama, the President at the time) whereas, in one experiment, the other was a friend. Parelman et al. found that a classifier could distinguish between patterns of activation for the self-reference and other-reference processing in the mPFC, PCC, and temporal lobes. The researchers also implemented a ridge PCR-model to examine the organisation of self and other-related activation in the mPFC. The neural code for self-reference and other-reference processing was not represented as linear dorsal-ventral gradient, but rather as a distributed pattern within the mPFC. Here, the peak weights for other-reference were located both in the ventral and dorsal parts of the mPFC (z = -14, z = 24), whereas the peak weights for self-reference were located in the ventral part of the mPFC (z = -2).

Another MVPA approach that is often used in fMRI analysis is representational similarity analysis (RSA; Kriegskorte et al., 2008). In contrast to MVPA with a classifier, which focuses on distinguishing between patterns of activation for different conditions, RSA focuses on the relative similarity between conditions. When fMRI data are analysed with RSA, it would, like the MVPA with a classifier, examine patterns of activation in the neural data. In other words, RSA would test the relative similarity in the neural patterns of activation. An advantage of RSA is that it is not limited to including only one type of data in the same analysis. For example, it can

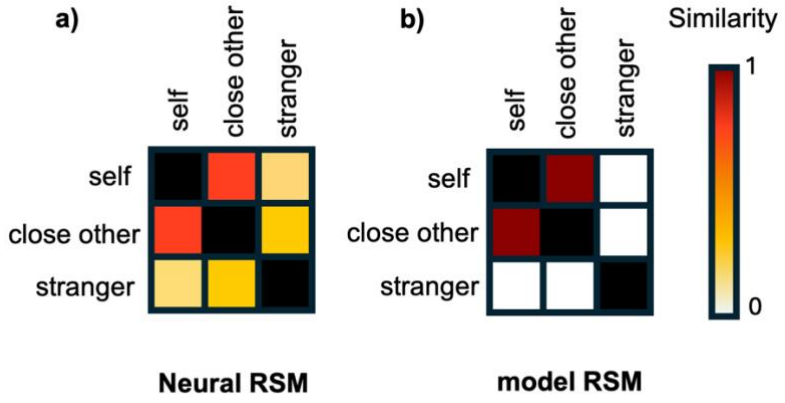
include both neural and behavioural data (e.g., fMRI data and questionnaire data), data from different neuroimaging methods (e.g., direct neural recordings and fMRI data), or even data from different species (e.g., mice and humans; Dimsdale-Zucker & Ranganath, 2019; Kriegeskorte et al., 2008). Although RSA can be used to compare categorical conditions (like in MVPA with a classifier, as explained above), it is also suitable for testing relationships with continuous variables (e.g., rating of mental states). That is, RSA can furnish answers to a variety of hypotheses. The benefit of including multiple types of data and also allowing for testing complex questions gives RSA a sizeable benefit in terms of flexibility compared to other types of analyses (Dimsdale-Zucker & Ranganath, 2019; Popal et al., 2020).

When performing RSA, the first step is to quantify similarity by creating a neural representational similarity matrix (RSM). An RSM is a symmetrical matrix that displays the similarity between pairs of conditions (or stimuli within condition) in an experiment. In the RSM, all conditions are displayed on both the vertical and horizontal side of the matrix. To create an RSM for an fMRI experiment, researchers would first extract the value from each voxel in the fMRI data (within a given spatial scale, see Figure 1.5). Then, they would calculate correlations between the values extracted from the fMRI data for pairs of conditions using Person's correlation, Euclidean distance, Mahalanobis distance or, cosine distances (Dimsdale-Zucker & Ranganath, 2019). Next, researchers would create a model RSM, the reference model that represents the hypothesised model. The model RSM can be computational (e.g., simulating a cognitive process) or behavioural (e.g., self-reference rating) testing the fit between the behavioural and neural data. Furthermore, the model RSM can be conceptual, based on an idea that researchers intend to test without specifying the computation or data on which it is based (Kriegeskorte et al., 2008). The next step would be to compare the model RSM to the neural RSM to find out how well the theoretical model represents the neural data. To test the fit, one would calculate the similarity (e.g., correlation) between the two models. This is often done by calculating Kendall's tau between the values in the two RSMs (Dimsdale-Zucker & Ranganath, 2019; Nili et al., 2014). Given that the RSMs are symmetrical, only the values in the lower (or

upper) triangle are entered into the calculation. The similarity between model RSMs and neural RSMs can also be visualised through heatmaps. It is also possible to carry out an RSA regression with multiple model RSMs to test which model RSM best predicts the neural RSM. The procedure for conducting RSA is completed separately for each participant before second-level analysis that tests if there is significant similarity across participants in the given region.

As an example, in Figure 1.6, RSA depicts a neural RSM of the similarity between patterns of activations for the different conditions in an experiment that compares self-reference to other reference for both a close other and a stranger. It also depicts a (conceptual) model RSM that represents similar patterns of activation for self and close other.

Figure 1.6 RSA Example



Also, experiments using RSA indicate that self-reference and other-references processing are based on different neural patterns of activation in the mPFC. For example, Feng et al., (2018) asked participants (*N* = 69) to complete the self-reference task and the other-reference task. In the other-reference task, participants made other-reference judgements of one's mother and a celebrity. Each condition presented words on three different dimensions: traits, physical attributes, and social roles. Neural responses in the mPFC (specifically in the vmPFC [z = -6]) and PCC distinguished the self from both the other two conditions (celebrity and mother). Also,

different dimensions of the self were distinguishable in the patterns of activation in these regions.

Similarly, Courtney and Meyer (2020) employed RSA to compare neural responses during self-reference and other-reference processing (N = 41). Here, the other-reference task included close others, acquaintances, and celebrities. In the dmPFC (peaks around z = 33, z = 45 and z = 42), the neural representation of the self was different from the representation of each of the other conditions, whereas close others and acquaintances were similarly represented to each other. Further, celebrities were represented by different patterns of activation compared to the self, close others, and acquaintances. In addition, the higher the subjective closeness, the higher was the similarity of neural patterns in the mPFC, PCC, and precuneus. These results suggest that (1) self-reference is represented separately from other-reference, and (2) representation in mPFC is distinguishable for various others, as well as different dimension of self and other.

Overall, experiments using an MVPA approach have demonstrated that the neural processing of self-reference and other reference are distinguishable in the mPFC when comparing the self with another person. Specifically, patterns of activation are distinguishable when comparing (1) the self to a close other, (2) the self to a minimally related other, and (3) the self to a stranger. The specific location of patterns of activation within the mPFC that distinguishes between the self and other varies across experiments. I elaborate on this issue next.

1.5.3 Spatial Organisation of Self and Other Activation Within the mPFC

The mPFC is a relatively large brain region, and, although self-reference and other-reference tasks both activate it (Wagner et al., 2012), how this activation is spatially organised remains a topic of debate. Some experiments suggest that self-reference and other-reference processes activate separate parts of the mPFC, with self-reference processes being localised in the vmPFC (van der Meer et al., 2010) and other-reference processes in the dmPFC (Van

Overwalle 2009). Other experiments, however, suggest the opposite (Koster-Hale et al., 2017; Lou et al., 2004; Schmitz et al., 2004; Seger et al., 2004). A meta-analysis by Denny et al. (2012) incorporated peak activation points from 107 experiments using the self-reference or other-reference tasks, and included the z-coordinates of the experiments in a logistic regression. Self-reference and other-reference activation was organised along a dorsal-ventral gradient within the mPFC. These results suggest that, although there is not a stark dorsal-ventral division within the mPFC, self-reference activation peaks are often located in the ventral part, whereas other-reference peaks are more commonly found in the dorsal part of the mPFC.

Yet, an electrocorticography (ECoG) experiment yielded contradictory results. Tan et al. (2022) used ECoG, a method that measures brain activation directly from populations of neurons, and thus can capture the fast spikes of neurons, while having high spatial resolution. The researchers recorded activation from several cortical regions: visual cortex, anterior temporal lobe, temporoparietal junction (TPJ), posteromedial cortex, and mPFC (ventral, dorsal and anterior parts). Participants (N = 16) were patients who had electrodes surgically implanted to their cortex to monitor and treat epilepsy prior to surgery. Participants carried out the self-reference task and the other-reference task (where the other was a neighbour). Activation for both the self- and other-reference task followed a common pathway which started in visual cortex, then temporoparietal regions, and finally the mPFC. Also, compared to self-reference, other-reference activations were slower and longer in all regions except the TPJ. In contrast to the abovementioned experiments and meta-analysis (Denny et al., 2012; Van der Meer et al., 2010; Van Overwalle et al., 2009), which suggest a spatial dissociation between self and other activation within the mPFC, Tan et al. found that (1) activation for both conditions recruited almost identical sites in the mPFC, and (2) the difference between the two conditions was only temporal, that is, other-reference processing started later and lasted longer. The authors speculated that the temporal difference could be due to the rich and more accessible information people have about themselves. Although this experiment only included 16 participants, EcoG is a powerful method. By directly recording brain activation at the level of

populations of neurons, Tan et al. obtained a relatively strong anatomical precision (Parvizi & Kastner, 2018) compared to the previous fMRI experiments (Denny et al., 2012; Van der Meer et al., 2010; Van Overwalle et al., 2009), that suggested a spatial dissociation of activation for self and other in mPFC.

In summary, early fMRI experiments yielded a spatial dissociation within the mPFC for neural responses of the self-reference and other-reference task, suggesting that neural activation for the self is located in the ventral portion, whereas neural activation for another person is located in the dorsal part. A meta-analysis expanded on these findings, proposing a linear dorsal-ventral gradient organisation in the activation patterns of the two tasks. Recently, an ECoG experiment challenged these findings, indicating that self- and other-related responses occur in nearly identical locations within the mPFC. So far, the majority of the outlined research has focused on comparing self-reference with other-reference, however, research on the self-reference task has also been investigated in relation to various other tasks. The following section will outline these studies and present potential explanations for why the mPFC is active during the self-reference task.

1.6 Explanations for mPFC Activation During Self-Reference

1.6.1 Self-Relevance Is Associated with mPFC Activation

In a few experiments using the self-reference task (Fossati et al., 2003; Macrae et al., 2004), mPFC activation was greater when participants responded “Yes, this word describes me” compared to “No, this word does not describe me.” Based on this finding, Moran et al. (2006) hypothesised that mPFC activation is modulated by the self-relevance of a stimuli. They also noted that self-relevance is often confounded with stimulus valence (i.e., emotional response to stimuli one considers self-relevant) and therefore controlled for this factor in their fMRI experiment. They asked participants (N = 42) to rate the self-descriptiveness (1 = *not at all like me*, 4 = *very much like me*) of both positive (e.g., honest) and negative (e.g., lazy) personality traits. They used parametric modulation analysis to examine the relationship between neural

activation and behavioural ratings of the stimuli. Activation in the mPFC and PCC linearly increased as a function of self-relevance. In addition, regardless of valence, the mPFC and PCC responded preferentially to self-relevant stimuli. D'Argembeau et al. (2012) replicated the findings of a positive relationship between activation in the mPFC and PCC and self-descriptiveness rating, and showed that vmPFC activation is positively associated with self-importance ratings.

Moran et al. (2009) compared self-reference judgements to social desirability judgements. They tested whether self-relevance ratings were also related to mPFC activation when participants (*N* = 30) made other types of ratings (i.e., "How socially desirable is this trait?") about the same stimuli, and not just during explicit self-descriptiveness ratings (i.e., "How much does this trait describe me?"). The authors considered the social desirability rating implicit because, in the task, participants made judgments and processed the stimuli without directly relating the items to themselves. The researchers replicated their prior finding (Moran et al., 2006) that the higher the rating, the greater the mPFC activation during the self-reference task. However, they obtained these results only for the explicit judgements task, and not the implicit (i.e., social desirability) task. It appeared that self-reflection is necessary for modulating mPFC activation based on the self-relevance of stimuli rather than merely responding to stimuli in a non-self-related task.

Phan et al. (2004) reported a positive association between self-relevance ratings and mPFC activation. Participants (*N* =12) rated the emotional intensity and self-relatedness of affective pictures while in the scanner. The higher the self-relatedness rating, the greater the activation in mPFC during the self-relatedness task was, whereas the emotional intensity rating was positively correlated with activation in the amygdala.

In an experiment by Northoff et al. (2009), participants (*N* = 15) passively viewed pictures while in an fMRI scanner, and then completed a self-relatedness rating, an emotional valence (i.e., pleasantness) rating, and an intensity rating. The higher the self-relatedness rating, the

greater the activation in the dmPFC, nucleus accumbens, right amygdala, tectum, thalamus, and hypothalamus and PCC was. Next, the researchers examined the association between activation and emotional valence, as well as between activation and intensity ratings, in regions where activation was correlated with self-relatedness. They found a negative association between emotional valence and activation in the nucleus accumbens and dmPFC. They also found a negative association between intensity ratings and activation in the nucleus accumbens and dmPFC. Finally, they found a positive association between emotional valence and activation in the amygdala and tectum, as well as between intensity ratings and activation in these regions. These results, along those of Phan et al. (2004), indicate that self-relevance is related to mPFC activation not only during self-related judgements of words but also when evaluating picture stimuli. The results additionally indicate that this association persists regardless of whether the stimuli are passively viewed or evaluatively judged.

Relatedly, the abovementioned study by Koski et al. (2020) (outlined in the "MVPA studies comparing self and other" section) also conducted an analysis to examine mPFC activation and self-relevance. When compared to irrelevant traits, they found that mPFC activation was greater for traits that were relevant to the self and a minimally related other, but not to an unrelated other.

Taken together, these six experiments (D'Argembeau et al., 2012; Koski et al., 2020; Moran et al., 2006, 2009; Northoff et al., 2009; Phan et al., 2004) showed a positive relationship between mPFC activation during self-related judgements and degree of self-relevance of the stimuli. However, the experiments only compared self-relevance judgements to desirability or emotional valence judgements or a minimally related other, but did not compare self-relevance judgments to those of a close other. Although many experiments have examined the link between the mPFC and the self-reference task (Wagner et al. 2012), and could easily have conducted and reported parametric modulation analysis, only these six experiments mentioned having conducted the analysis and found a linear relationship between mPFC activation and self-relevance. Unreported null findings on the association between mPFC activation and self-

relevance ratings are possible. So, due to the absence of experiments reporting similar findings, along with the lack of a "close other" control condition, it is unclear whether the mPFC specifically processes the self-relevance of stimuli. Another potential explanation for mPFC activation during the self-reference task is that the mPFC is responding to the value of stimuli rather than their self-relevance. In the following section, I review literature on the self in relation to value.

1.6.2 The Value Hypothesis: Self-Reference versus Reward

Given that the vmPFC was found to be active during the self-reference task and while processing value (Delgado et al., 2016), it was suggested that its activation in the self-reference task reflects the tracking of stimulus value (D'Argembeau, 2013). This idea aligns with Krienen et al.'s (2010) conclusion that vmPFC activation is greater for close others than strangers, because the vmPFC tracks the social value of others. D'Argembeau (2013) proposed that mPFC activation during the self-reference task reflects the process of assigning value or significance of self-related stimuli. This, so called "value hypothesis" received preliminary support in experiments showing that mPFC is active during the processing of self-related information such as traits, attitudes, values, physical attributes, goals, memories, future thoughts, close others, social groups, and possessions (Denny et al., 2012; Martinelli et al., 2013; Murray et al., 2012; van der Meer et al., 2010).

D'Argembeau et al. (2012) tested the value hypothesis. While in a scanner, participants (*N* = 23) performed the self-reference task. After the scan, they rated the self-descriptiveness (i.e., "to what extent does the trait describe you?"), certainty (i.e., "how certain are you that you possess or do not possess this trait?"), and self-importance of each trait (i.e., "how important is it for you to possess or not possess this trait?"). The researchers found a positive correlation between activation in the vmPFC, dmPFC, and PCC during the self-reference task and the self-descriptiveness rating. They also observed a positive correlation between dmPFC activation during the self-reference task and the certainty rating. Furthermore, they reported a positive

correlation between activation in the vmPFC and self-importance ratings. Because participants made no explicit value judgement during the scan, the researchers speculated that the mPFC automatically assigns value to a trait. These results are consistent with those of other neuroimaging experiments, suggesting that the mPFC is involved in processing the self-importance of stimuli (D'Argembeau et al., 2010; Schmitz & Johnson, 2007).

Researchers have also found an overlap in mPFC activation for self-reference and reward processing of emotional stimuli. Enzi et al. (2009) compared neural activation during judgements of personal relevance and reward processing. Participants (*N* = 19) performed three tasks. In the reward task, they pressed one of two buttons to decide on a gambling choice. Subsequently, they were presented with the outcome, either a win or a loss of money. The total sum across trials would determine their reimbursement. In the self-relevance task, participants viewed pictures and judged if these were high or low on self-relevance. In the control task, they judged if pictures were vertically or horizontally aligned. When Enzi et al. compared win trials to lose trials and high relevance to low relevance, they found an overlap in activation in the pregenual cingulate cortex, caudate nucleus, ventrolateral prefrontal cortex, putamen, insula, and dmPFC, indicating that these regions are active during both reward and personal relevance processing. They also found activation in the anterior insula, supragenual ACC, and premotor cortex when comparing high personal relevance to low personal relevance. They concluded that their findings are compatible with the value hypothesis, as the overlapping regions processed both immediate rewards (i.e., money) and long-term rewards (i.e., self-relevant information).

De Greck et al. (2008) also compared activation during the processing of self and reward. Participants (*N* = 15) engaged in three tasks: gambling, self-relatedness, control. In all tasks, they viewed stimuli related to food, alcohol, and gambling. In the gambling task, participants imagined that they were gambling about the picture presented. They had to make a gambling decision, that is, to bet on the right or left side. In the self-relatedness task, they judged if the picture was self-related or not. In the control task, they judged if the stimuli were vertically or horizontally aligned. Next, participants were presented with an outcome. In the gambling and

control tasks, they viewed a "+" sign for win or correct trial and a "-" lose or incorrect trials. In the self-relatedness task, they viewed an "=" sign regardless of their response. De Greck et al. first compared activation during the win trials to the lose trials in the gambling task. They found activation in the brain's reward network including the nucleus accumbens, vmPFC, and ventral tegmental area. Subsequently, they compared the signal change in these three regions during win and lose trials as well as self-high and self-low trials. They found a higher signal in all three regions when they compared win with lose and self-high with self-low. Similarly to Enzi et al. (2009), they concluded that self-relatedness and reward are processed in these three regions, in support of the value hypothesis.

The findings by Enzi et al. (2009) and De Greck et al. (2008) are also in line with previous research such as those of Phan et al. (2004) (outlined above), who reported that vmPFC tracks the self-relevance of emotional pictures. Based on the activation overlap for self and reward in the vmPFC, Northoff and Hayes (2011) questioned whether self-reference processing is just a type of reward processing. However, univariate analyses, which these abovementioned experiments used, cannot offer conclusive evidence on whether reward and self-reference share a common neural mechanism.

Chavez et al. (2017) employed Multivariate cross-classification (MVCC; Kaplan, 2015) to investigate the neural basis of positive affect and self-reference. MVCC is an alternative way of using MVPA with a classifier, just that here the test-data are from two conditions and the training-data are from two separate conditions. MVCC can clarify whether a classifier can generalise from one cognitive context to another. For example, researchers could employ this approach to examine if viewing positive valence photographs and the self-reference task are based on common neural patterns of activation. Here, researchers would first train the classifier to distinguish between the self-reference task and a control task, such as the other-reference task. They would train the classifier on all data (runs) for these two tasks. They would next apply the classifier to test if it can distinguish between the positive valence task and the negative valence task (control condition), using the ensuing data as test data (Figure 1.6b). If the

classifier can generalise, then there is learning transfer from training to testing, and the researcher can infer that the two tasks are based on a common neural mechanism.

Chavez et al. (2017; *N* = 14) used MVCC with a searchlight approach. First, they trained a classifier to distinguish between participants' patterns of activations while rating the valence of positive and negative images. When the researchers applied the classifier to patterns of activation during the self-reference versus other reference task, they found that the classifier could successfully distinguish the patterns of activation in the vmPFC. To examine if classification accuracy was due to similarity between self and positive affect or between other and negative affect, they also conducted RSA. When comparing the conditions, they found the highest similarity for the neural patterns of activation for self-reference and positive affect in the vmPFC. Thus, their results suggest that self-reference and positive affect are based on a similar neural mechanism. These results are consistent with the reward experiments (De Greck et al., 2008; Enzi et al., 2009; Phan et al., 2004) and the value hypothesis (D'Argembeau et al., 2013).

An experiment by Yankouskaya et al. (2017) yielded similar findings. In a shape-label matching task, participants (*N* = 16) completed two tasks, a self-task and a reward task. In both tasks, they imagined pairing two geometric shapes with two tags: self and other tags in the self-task (e.g., self-square, other-circle), low-reward and high-reward in terms of monetary amounts in the reward task (e.g., £1-hexagon, £16-triangle). While in an fMRI scanner, participants were presented with a different tag and shape on each trial. They judged if each combination was a match or mismatch from the initial associations. Next, they applied a classifier trained on the neural data for one task (e.g., self versus other) to the other task (e.g., high versus low reward). The results indicated that a classifier could successfully generalise self to reward in the vmPFC. The ROIs with above chance classification accuracy for the two experiments, overlapped in an anterior part of the vmPFC (Broadmann area 10).

In summary, as the vmPFC is active during the processing of value and stimuli that would be significant to oneself and also during the self-reference task, a key reason for vmPFC

activation during the self-reference task may be to track the value of stimuli. Experiments that compared the activation during self and reward processing found an activation overlap in the vmPFC. However, due to limitations of univariate analyses, it was not until MVPA experiments began to examine activation during the two tasks that research could identify a common neural code for self-reference and positive affect/reward in the vmPFC. Autobiographical memory may also explain mPFC activation during the self-reference task. Next, I address this topic.

1.6.3 Self-Reference versus Autobiographical Memory

Autobiographical memory tasks activate the mPFC (Summerfield et al., 2009). It is unclear, though, whether processing of autobiographical memories takes place during the self-reference task. Some fMRI experiments suggest that autobiographical memory retrieval is involved in self-reference processing (Araujo et al., 2014; Kelley et al., 2002). This suggestion, though, runs contrary to relevant psychological models (Conway 2005; Haslam et al., 2010; Klein, 2010) and neuropsychological findings (Klein & Lax, 2010; Tulving et al., 1993).

Case studies of patients with memory impairments as a result of neural damage point to the resilience of self-trait knowledge in memory (Klein & Lax, 2010). Tulving (1993) examined different types of self-related memory in patient K.C. who suffered from amnesia due to an accident. K.C. was unable to recollect any episodic memory. However, when asked to rate the self-descriptiveness of adjectives, his ratings were both reliable over time and consistent with ratings of a close other. The patient, then, had intact knowledge of his own personality, although he had no conscious access to episodes. Other studies with amnesic (Klein et al., 1996, 2002) and Alzheimer's (Klein et al., 2003) patients have reported similar findings. Patient D.B., who suffered from amnesia, showed intact self-trait knowledge and was unable to retrieve episodic memory (Klein et al., 2002). However, his trait knowledge of a close other was impaired. Specifically, he rated his daughter's personality on two occasions separated by a week. His daughter rated herself on the same personality traits. D.B.'s rating of his daughter and her rating of herself was uncorrelated (r = .23), and although D.B.'s ratings of his daughter were somewhat

reliable across the two testing sessions (r = .58), they were lower in reliability than those of control participants (r = .90). Taken together, whereas some memory systems are dysfunctional (e.g., episodic memory, trait knowledge of others), self-trait knowledge is preserved and likely stored separately from episodic memory.

Three meta-analyses of 38 experiments in total, reported by Martinelli et al. (2013), converged to the conclusion that episodic memory and self-trait knowledge are stored separately. In particular, the authors tested a model of autobiographical memory that divides autobiographical memory into three related, but functionally independent, systems: (1) episodic memory, that is, memory of specific episodic events; (2) semantic memory, that is, memory of general knowledge of personal facts; (3) self-knowledge of personality traits (Conway 2005). Episodic memory mainly recruited posterior and limbic regions. Semantic memory recruited anterior regions as well as posterior and limbic regions but to a less extent than episodic memory. Finally, self-knowledge of personality mainly recruited the mPFC. These results resonate with the idea that, during self-reference judgements of traits, the mPFC accesses pre-computed summaries of traits rather than semantic or episodic memories. The results bolster the findings of previous case studies showing that self-trait knowledge is stored separately from episodic memory (Klein & Lax, 2010; Klein et al., 1996, 2002, 2003; Tulving, 1993).

A recent ECoG experiment tested whether neural responses during self-reference and autobiographical memory processing overlapped (Iravani et al., 2024). Participants (N = 22) were epilepsy patients with implanted ECoG recording sites, waiting to undergo surgery. They completed three tasks: (1) a self-reference task where they judged if a trait describes them; (2) an episodic memory task where they judged if a statement was true (e.g., “I went on a walk today”); (3) an arithmetic task where they judged if calculations were correct (e.g., 47 + 8 = 55). Neural recordings were implanted in the orbitofrontal cortex (oPFC) and vmPFC. The results indicated that neural responses were overall sparse for both the self-reference and autobiographical memory tasks, where only 33 of 253 recording sites showed greater activation

during the two tasks compared to the arithmetic task. Also, the majority of the neural responses for the self-reference and autobiographical memory tasks occurred in the oPFC and, importantly, did not overlap. These findings are consistent with the abovementioned meta-analyses and case studies, pointing to separate processing of self-reference and autobiographical memory. Furthermore, for both the self-reference and autobiographical memory task the recording occurred earlier in the vmPFC than in the oPFC, suggesting that processing of the stimuli happens in the vmPFC first. Although ECoG offers exceptional spatial and temporal precision, and the opportunity to directly record from neural populations in humans remains rare, this experiment has limitations. The measures were restricted to the implanted sites and therefore included the oPFC and vmPFC, but not dorsal, parts of the mPFC. Moreover, the autobiographical memory task only incorporated the retrieval of very recent, as opposed to long-term, episodes.

In summary, although some early fMRI experiments considered autobiographical memory as a part of self-reference (Araujo et al., 2014; Kelley et al., 2002), case studies, meta-analytic evidence, and an ECoG experiment indicate that autobiographical memory and self-reference trait processing are functionally and anatomically separate in the brain (Iravani et al., 2024; Klein & Lax, 2010; Martinelli et al., 2013). I consider next literature on self-reference in relation to introspection.

1.6.4 Self-Reference versus Introspection

Another process which activates the mPFC and that might be related to thinking about the self is introspection. Introspection refers to thinking about one's own mental states (Seitz et al., 2009). Goldberg et al. (2006) asked participants (N = 9) to complete five tasks while in an fMRI scanner. In the introspection task, participants thought about how they felt while viewing both visual and auditory stimuli. The visual stimuli consisted of positive and negative images, whereas the auditory stimuli consisted of musical clips. Participants judged whether the stimuli made them feel emotionally aroused (positively or negatively) or neutral. In the categorisation

task, participants viewed the same stimuli but judged if there were animals or no animals in the visual task and whether there were trumpets or no trumpets in the auditory clip. They also complete a semantic judgement task, judging if a word was a verb or a noun, as well as a picture judgement task, judging if a word was positive or negative, and the self-reference task. Goldberg et al. reported activation in the mPFC, when comparing the self-reference task to the semantic condition. They also reported activation in the mPFC, when comparing the introspection task to the categorisation task. Furthermore, activation in the mPFC from these two contrasts (self-reference versus semantic and introspection versus categorisation) overlapped. Finally, the auditory task activated the mPFC (i.e., ACC and superior frontal gyrus).

Relatedly, in an fMRI experiment, Araujo et al. (2015; *N* = 19) examined interoception in relation to self-reference. Interoception refers to perceiving one's own internal bodily states (Schulz et al., 2016). In the study, Araujo et al. compared interoception (e.g., "Do you feel hungry?") to exteroception (e.g., "Do your legs feel wet?"), and self-reference judgements of biographic facts (e.g., "are you a student?") to traits ("e.g., "Does the word "honest describe you?"). They found greater mPFC activation in a ventral part of the mPFC, when comparing the self-reference task of traits to that of biographical facts. They also found greater mPFC activation in a ventral part of the mPFC, when comparing biographical facts to traits. Likewise, Araujo et al. found greater activation in the mPFC, when comparing both self-reference conditions (facts and traits) to baseline. This was also the case when comparing the two self-reference tasks to both the interoception and exteroception tasks. Lastly, the researchers found greater activation in the mPFC during the interoception task, when comparing this task to the exteroception task.

In summary, the results of these experiments suggest that introspection and self-reference both activate the mPFC. However, the relevant sample sizes were small. Additionally, as mentioned above, activation overlap in univariate experiments does not indicate a common neural mechanism. Therefore, it continued to be unclear whether common mPFC activation during self-reference and introspection tasks reflects common cognitive processes. Further,

other internal mentation processes can activate the mPFC along with a network of other regions, the default mode network (DMN). I review next literature on the self and the DMN.

1.7 The Self and the Default Mode Network

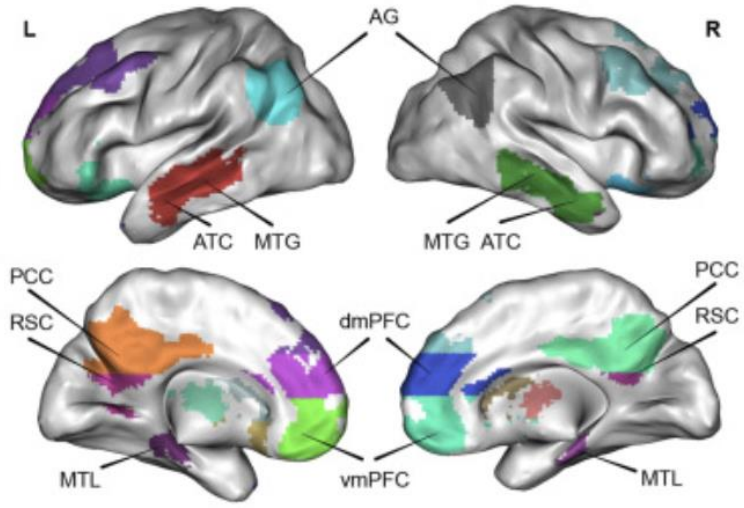
In the late 90s, researchers discovered a network of brain regions that showed greater activation during passive viewing than during memory and goal-directed attention tasks (Shulman et al., 1997). They labelled this network the DMN (Raichle et al., 2001). It consists of the PCC, mPFC, retrosplenial cortex, anterior temporal cortex, middle temporal gyrus, medial temporal lobe and angular gyrus, anterior and mediodorsal thalamic nuclei, as well as septal nuclei and nucleus accumbens (Alves et al., 2019) (Figure 1.7).

Initial findings sparked numerous research questions, including whether the DMN was a coherent network. Greicius et al. (2003) examined the functional connectivity between regions that showed suppressed activation during memory tasks. Participants (N = 14) performed three tasks: (1) a working memory task where they were asked to remember the spatial location of a circle; (2) a passive visual task where they viewed images of checkerboards; (3) a resting state task where they closed their eyes and did not think of anything specific. Greicius et al. used the working memory task to identify regions that showed a decrease in activation during an active task. The researchers compared activation during the working memory task to a baseline control task that involved pressing a button if a circle on the screen was in a central location. They found a decrease in activation in the PCC and vACC. Next, they examined the functional connectivity between these two regions and other regions that were active during the resting state task and the passive visual task. Functional connectivity measures the correlation of fluctuations in activation across different regions over time. Here, a high correlation between regions is assumed to reflect a tightly connected network. Greicius et al. reported that, during both the resting state task and the passive visual task, the ventral ACC was functionally connected to the mPFC, midbrain, nucleus accumbens, and PCC. Furthermore, PCC was functionally connected to the mPFC, dorsolateral prefrontal cortex, inferolateral temporal

cortex, and parahippomcampal gyrus, during both tasks. These findings were the first to show that the DMN was indeed an interconnected network during resting state.

Early research linked the DMN to thinking about the self (Wicker et al., 2003). Resting state was often associated with mental processes related to a relaxed state such as mind wandering, day dreaming, and stimulus-independent thoughts (Raichle, 2015). These spontaneous thoughts were often focused on information related to the self. For example, Andrews-Hanna et al. (2010) found that spontaneous thoughts usually referred to self-relevant issues. Baird et al. (2011) found that during a reaction time task, task unrelated thoughts consisted of thoughts regarding the self 66% of the time. Stawarczyk et al. (2011) examined the subjective reports of thought content during an attention task. They observed that participants were frequently focused on their personal goals. Although early research on the DMN was primarily concerned with resting state and spontaneous thoughts, later research discovered that the DMN is also active during some tasks that require internal mentation. In particular, the DMN is involved in directed self-related processes such as self-reference (Davey et al., 2016). Further, the DMN is involved in processing social cognition (Molenberghs et al., 2016), episodic memory (Spreng et al., 2009), semantic memory, and language (Binder et al., 2009; Humphreys et al., 2021).

Figure 1.7 Cortical Regions Included in the Default Mode Network (Menon et al., 2023)



Prompted by findings that resting state includes thinking about the self and activates similar regions to self-reference (Kelley et al., 2002), researchers began to compare the two processes (D'Argembeau et al., 2005; Qin & Northoff, 2011). In a PET experiment, D'Argembeau et al. (2005) asked participants (*N* = 12) to reflect on their own personality, a friend's personality, a social issue, and to not think in a systematic way (resting state). After the scan, participants described and rated what they had thought of during each condition in the scan. The vmPFC showed greater activation during the self-reflection task compared to reflecting on the friend or a social issue. Conjunction analysis indicated an activation overlap in the vmPFC for the self-reflection task and resting state condition. D'Argembeau et al. also found a positive within-subjects correlation between vmPFC activation and the amount of time spent thinking of the self across all conditions. Moreover, participants reported to have spent more time thinking about the self in the self-reflection condition compared to rest, but to have spent more time thinking about the self in the rest condition compared to the other-reflection condition. D'Argembeau et al. concluded that, during resting state, people likely have a variety of thoughts including self-related ones. Based on the activation overlap, they further concluded that, during resting state, people think about information related to themselves, facilitating a stable and coherent self-concept.

Building on findings that mPFC is active during self-reference processing and is also a key part of the DMN, which remains relatively active at rest, Meyer and Lieberman (2018) conducted

an experiment to examine default self-thinking. Previous experiments on self-processing and the DMN had suggested that, during rest, people think about themselves, which in turn activates the mPFC. However, Meyer and Lieberman proposed the reverse: mPFC becomes by default active at rest, triggering self-focused thinking. These researchers wondered whether mPFC activation at rest primes self-reference thinking. Participants (N = 19) completed three tasks, while in an fMRI scanner: (1) self-reference task; (2) other-reference task, judging if traits described Obama (the U.S. President at the time); (3) location judgement task, judging if a trait described the Grand Canyon. Prior to each trial, participants went through a resting phase (a cross on the screen for 6-9 seconds). After each task trial, they completed an attention orienting trial, intended to reset their thinking. Meyer and Lieberman reasoned that, if mPFC activation at rest primes people to think about themselves, then greater magnitude of mPFC activation in the rest phase will facilitate reaction times in the following trial, provided that the following trial involves self-reference, but not in the other two tasks. They focused on three mPFC ROIs (reverse inference maps derived from Neurosynth by searching for "self", "social cognition", and "semantic"). As hypothesised, mPFC activation during rest primed subsequent self-reference. Specifically, activation in the anterior mPFC (Broadmann's area 10) primed self-reference, whereas activation in the dmPFC (Broadmann's area 9) primed both self-reference and other-reference. They concluded that, at rest, the mPFC activates by default, triggering self-thinking, which explains why people often revert to thinking about themselves.

A range of explanations have been put forward to explain the overall function of the DMN. D'Argembeau et al. (2010, 2018) proposed that the DMN processes self-relevance. Specifically, the DMN integrates various representations to form self-referential thoughts, thus creating a sense of identity and facilitating planning or goal pursuit. Given that the DMN is recruited for tasks including envisioning the future, remembering the past, navigation, and taking others' perspective, Buckner and Carroll (2007) posited that the DMN's function is to use past experiences for imagining perspectives and events independently of the immediate environment. Menon (2023) argued that the DMN integrates various cognitive processes to

construct an internal narrative, which is important for one's sense of self and shapes understanding of personal experiences. Other researchers, however, maintained that the DMN's function is not directly related to processing information about the self. For example, according to Hassabis and Maguire (2007), the purpose of the DMN is scene construction. According to Konishi et al. (2015), the DMN's role is to shape cognition by internal representations that are independent of immediate perceptual input. And according to Yeshurun et al. (2021), the DMN is a sense-making network that is active and dynamic: it integrates extrinsic and intrinsic information to make context-dependent models as they unfold over time.

The DMN literature has had its challenges. First, it is difficult to investigate some aspects of the human DMN by using animal models. In neuroscience, understanding the roles of brain networks in human cognition often involves comparing them to animals. Although, some experiments have identified an equivalent version of the DMN in monkeys and rats during resting state (Lu et al., 2012; Stafford et al., 2014; Vincent et al., 2007), autobiographical memory, self-reference, and mind-wandering are processes that cannot be measured in animals (Menon, 2023), making it very challenging to investigate the roles of the DMN with animals. Secondly, whereas some research has attempted to understand how subnetworks of the DMN are interconnected (Andrews-Hanna et al., 2010), the DMN is involved in processes that have been studied in isolation. For example, the literatures on the DMN's role in language processing, self-reference, and autobiographical memory have progressed separately for the most part. Thus, the integration of these cognitive processes and the DMN's overall function remains unclear.

Taken together, the DMN is a network of brain regions involved in both spontaneous self-related thoughts that occur independently of external demands and tasks that requires internal mentation, including the self-reference task (D'Argembeau et al., 2005, 2018). Relevant findings indicate that people may be prone to routinely think about themselves when their minds are free of external demands by default, as a result of mPFC activation (Meyer & Lieberman, 2018).

Furthermore, the overall function of the DMN could be to create an internal narrative that facilitates one's sense of self (Menon, 2023).

1.8 Summary

The neural representation of the self has been studied in regard to other-reference, autobiographical memory, and introspection as well as social-relevance, value, and the DMN. Various methods such as fMRI, neuropsychology, ECoG, and analytic approaches like univariate analysis, MVPA, and functional connectivity have been used (Andrews-Hanna et al., 2010; D'Argembeau et al., 2013; Jenkins et al., 2008; Marquine et al., 2016, 2013; Phan et al., 2004; Tan et al., 2022; Wagner et al., 2012, 2019).

Research on the self in the brain has taken several stands. First, psychologist have proposed an associative network model of how self-knowledge is represented in memory (Anderson et al., 1976; Greenwald et al., 2002; Klein et al., 1989). Second, behavioural experiments have identified enhanced recall of stimuli judged in relation to the self—a self-reference effect (Rogers et al., 1977; Symons & Johnson, 1997). Third, a link has been discovered between the mPFC and the self-reference task (Murray et al, 2012). Fourth, multivariate experiments have indicated that the self and other are processed by different patterns of activation in the mPFC (Courtney & Meyer, 2020; Feng et al., 2018; Koski et al., 2020; Parelman et al., 2022), but also seemingly at identical locations in the mPFC (Tan et al., 2022). Fifth, explanations proposed to account for mPFC activation during the self-reference task include self-relevance (Phan et al., 2004), reward (De Greck et al., 2008), autobiographical memory (Martinelli et al., 2013), and introspection (Goldberg et al., 2006). Findings point to a common neural code for self-reference and positive affect/reward in the vmPFC (Chavez et al., 2017; Yankouskaya et al., 2017). Sixth, both spontaneous and directed self-related thought is processed by the DMN (Andrews-Hanna et al., 2010; D'Argembeau et al., 2005), a network of brain regions that shows greater activation during rest and internal mentation compared to externally demanding tasks (Menon et al, 2023).

This literature has limitations. To begin, research comparing self and close other is likely confounded by personal familiarity and previous knowledge (Gillian & Farah, 2005; Koski et al., 2020). Also, mPFC activation during the self-reference task could be a result of general cognitive processing taking place during the task rather than any self-specific process (Legrand & Ruby, 2009). Although, in recent years, MVPA experiments (Courtney & Meyer, 2020; Feng et al., 2018; Koski et al., 2020; Parelman et al., 2022) have overcome some of these limitations, there are still lingering questions about the representation of the self in the brain.

1.9 Thesis Outline and Aim

I aim to further knowledge about the self's representation in the brain. I control in the reported experiments for potential confound, investigating the multidimensional self-concept and using sensitive methods of analyses that address specific questions. I present three empirical chapters.

The first one (Chapter 2) consists of two fMRI experiments with a searchlight RSA approach to test where and how the self-concept is represented in the brain. I asked participants to provide items related to themselves. To measure the multidimensional self-concept, the items they provided could be anything responding to the prompt "I", which included social roles, likes, dislikes, and physical characteristics. Participants rated the items provided, in addition to some items provided by the experimenter, on self-descriptiveness and self-importance. While in the scanner, they performed the self-reference task and a control task (i.e., word-class judgement task in Experiment 1, other-reference task in Experiment 2) in which they were presented with stimuli they had previously rated. I tested the effect of self-importance and self-descriptiveness on neural patterns of activation, controlling for potential confounds such as valence, familiarity, and autobiographical memory.

The second empirical chapter (Chapter 3) consists of a behavioural experiment with a mixed-effect model analysis approach. Based on psychological theories of the structure of the self-concept reviewed above (Cantor & Kihlstrom, 1984; Greenwald et al., 2002), I tested the

psychological meaning of the associative links that connect self-related concepts to the self and to each other in an associate network model of the self. As the literature suggests that associative links can be measured in terms of reaction time facilitation, I used an evaluative priming task (EPT). Participants listed items that comprised their self-concept, and rated these items on self-importance, self-descriptiveness, and valence. I examine the influence of these three variables on reaction time. I hypothesised that the links in an associative memory model of the self represent self-importance.

The third empirical chapter (Chapter 4) consists of an fMRI experiment, featuring an MVPA and RSA searchlight approach. As mentioned above, various tasks have previously activated the mPFC, the hub of the DMN. Some authors have assumed that the mPFC is self-specific, although mPFC activation during the self-reference task could be due to general cognitive processes rather than anything related to the self per se. I tested similarities and differences in neural patterns of activation for the self-reference task and three other tasks, namely, other-reference, autobiographical memory, and introspection. Furthermore, by using variance portioning analysis, I investigated the extent to which mPFC activation patterns could be explained by those during the other three tasks. The final chapter (Chapter 5) contains a General Discussion.

Chapter 2 The Self-Concept is Represented in the Medial Prefrontal Cortex in Terms of Self-Importance[1]

2.1 Abstract

Knowledge about one's personality, the self-concept, shapes human experience. Social cognitive neuroscience has made strides addressing the question of where and how the self is represented in the brain. The answer, however, remains elusive. I conducted two functional magnetic resonance imaging experiments (the second preregistered) with human male and female participants employing a self-reference task with a broad range of attributes and carrying out a searchlight representational similarity analysis (RSA). The importance of attributes to self-identity was represented in the medial prefrontal cortex (mPFC), whereas mPFC activation was unrelated both to self-descriptiveness of attributes (Experiments 1 and 2) and importance of attributes to a friend's self-identity (Experiment 2). Our research provides a comprehensive answer to the abovementioned question: The self-concept is conceptualized in terms of self-importance and represented in the mPFC.

2.2 Introduction

The sense of self shapes human experience (Sedikides et al., 2021). The self (or self-concept) consists of knowledge that people possess about the kind of person they are, such as traits, physical attributes, preferences, beliefs, values, or ingroup (Sedikides & Gregg, 2003). The self has been of keen interest to psychologists since the birth of the discipline (James, 1890).

[1] This Chapter is based on the following published article: Levorsen, M., Aoki, R., Matsumoto, K., Sedikides, C., & Izuma, K. (2023). The self-concept is represented in the medial prefrontal cortex in terms of self-importance. *Journal of Neuroscience*, *43*(20), 3675-3686. https://doi.org/10.1523/JNEUROSCI.2178-22.2023

From the late 90s (Craik et al., 1999) onward, cognitive neuroscientists have been investigating the neural basis of the self (Wagner et al., 2019). However, where and how the self is represented in the brain remains elusive.

Past neuroimaging experiments on the self have identified a network of brain regions, including medial prefrontal cortex (mPFC) and posterior cingulate cortex (PCC), that are consistently active during self-reference judgment compared with other semantic judgements (Denny et al., 2012; Murray et al., 2012). However, the approach of contrasting neural responses during the self-reference versus control tasks to unveil the neural basis of the self has several limitations.

First, activation observed by simply comparing the strength of neural responses between the self-reference and control tasks may be because of cognitive processes unrelated to the self (Gillihan & Farah, 2005; Legrand & Ruby, 2009) such as autobiographical memory (Martinelli et al., 2013) and positive affect (Bartra et al., 2013). This limitation is at least partially addressed by recent functional magnetic resonance imaging (fMRI) experiments that used a multivariate pattern analysis (MVPA; Chavez et al., 2017; Courtney and Meyer, 2020; Feng et al., 2018; Koski et al., 2020; Parelman et al., 2022; Yankouskaya et al., 2017). Yet, these experiments have come short of documenting what information about the self is specifically being processed.

Second, exceptions notwithstanding (Rameson et al., 2010; Jenkins & Mitchell, 2011), the bulk of the literature has used only trait adjectives as experimental stimuli. However, this practice likely limits researchers' ability to identify the neural representations of the self. As stated above, the self includes not only personality traits, but also physical characteristics, preferences, aspirations, abilities, and social groups (Linville, 1985); thus, personality traits comprise a narrow subset of the self (see del Prado et al., 2007).

Third, most neuroimaging research has operationalized the self-concept in terms of trait self-descriptiveness. There is an infinite number of characteristics that can describe an individual (e.g., “I sleep everyday”), but just because an item is self-descriptive does not

necessarily mean it is a part of the self-concept. Instead, the self might be represented in the brain in terms of personal importance of each characteristic (hereafter, self-importance or centrality). The relevance of taking into account self-importance when assessing the self has also been recognised by psychologists (Markus, 1977). That is, whether information will influence one's behaviour depends on its personal importance (Markus, 1983). For example, if being a mother is important to an individual, her behaviours as a mother are likely to be different (e.g., more attentive, responsible, or consistent) from her behaviours in other, less self-defining roles (Deutsch et al., 1988). How self-important (central) a personality trait is affects how a person seeks (Sedikides, 1993) and remembers (Sedikides et al., 2016) information about themselves. Thus, it is likely that there is a dedicated neural system in the brain, which encodes the self-importance of incoming information (Markus & Wurf, 1987; Sedikides, 1995). Indeed, a few experiments have forayed into self-importance in the brain, suggesting that mPFC activation is correlated with the importance of possessing a personality trait (D'Argembeau et al., 2012) and with the personal significance of autobiographical memories (Lin et al., 2016).

I addressed the question of where and how the self-concept is represented in the brain in two fMRI experiments. I used the self-reference task with a broad range of stimuli combined with a representational similarity analysis (RSA; Kriegeskorte et al., 2008) of fMRI data to test how mPFC activation patterns are related to self-importance as well as self-descriptiveness, controlling for other factors (Materials and Methods).

2.3 Materials and Methods - Experiment 1

2.3.1 Participants

I recruited 32 right-handed undergraduate students from Tamagawa University, Japan. The students had no history of psychiatric disorders. I excluded data from four participants because of excessive head movement (> 3 mm; one participant), because their response consistency in the fMRI tasks was close to chance (one participant), because of no variance in the postscan memory rating (one participant), and because their self-reference rating reliability

was low (one participant). In regard to the last case, each participant completed the self-reference task three times, and I computed correlation coefficient across the three ratings. The average correlation of this fourth participant was 0.21, which was >3 SDs below the group average of r = 0.72 (SD = 0.14), suggesting very poor compliance with the task instructions and/or having a highly unstable self-concept. The final sample consisted of 28 participants (16 women, 12 men) aged 19–22 years (M = 19.84, SD = 0.86). Participants clicked a box to indicate their consent before the online questionnaires. The experiment was approved by both the University of Southampton and Tamagawa University ethics committees. I remunerated each participant with 5000 Japanese yen.

2.3.2 Experimental Procedure

The experiment comprised the three following parts, which took place on three separate days: (1) first online questionnaire; (2) second online questionnaire; (3) fMRI experiment. I administered the two online questionnaires in an effort to create the stimulus set for the fMRI experiment, a stimulus set that covers as widely as possible the content of each participant's self-concept (see below). The first and second online questionnaires were separated by an average of 6.25 days (SD = 3.26). The second online questionnaire and fMRI experiment were separated by an average of 8.29 days (SD = 2.19).

2.3.2.1 First Online Questionnaire

The first online questionnaire is similar to the Twenty Statement Test (TST; Kuhn & McPartland, 1954). During the online questionnaire, I instructed participants to provide at least 30 characteristics by responding to the prompt “I_____.” To facilitate this task, I gave participants examples such as physical characteristics (e.g., I am tall), personality (e.g., I am social), likes or dislikes (e.g., food, music, artists), and groups to which they belonged (university, department, clubs).

2.3.2.2 Second Online Questionnaire

The second online questionnaire included a total of 80 items some of which were provided by participants during the first online questionnaire, and others were added by an experimenter. Items prepared by the experimenter were intended to dissociate self-descriptiveness, self-importance, and other factors described below. For example, "right-handed" was added with an aim to dissociate self-descriptiveness and self-importance. "School trip" was added with an aim to dissociate self-descriptiveness/self-importance and autobiographical memory. "Convenience store" was added with an aim to dissociate self-descriptiveness/importance and familiarity. I instructed participants to rate each item in terms of: (1) self-descriptiveness (1 = *not descriptive at all*, 7 = *very descriptive*); (2) self-importance (1 = *not important at all*, 7 = *very important*); (3) valence (1 = very negative, 7 = *very positive*); (4) familiarity (1 = *not familiar at all*, 7 = *very familiar*); and (5) autobiographical memory or the extent to which "each word/phrase brought back memories of your past when you saw it" (1 = *it did not evoke any memory at all*, 7 = *it evoked very vivid memory*).

2.3.2.3 Stimulus Set Preparation

Based on the ratings, I selected a final stimulus set of 40 items under the stipulation that the ratings not be highly intercorrelated (i.e., effects on neural activities be dissociable). Specifically, I randomly picked 40 out of the 80 items used in the second questionnaire and computed correlations across the following six ratings/characteristics of the randomly selected 40 items: (1) self-descriptiveness; (2) valence; (3) familiarity; (4) autobiographical memory; (5) number of characters; (6) whether the item was provided by a participant during the first questionnaire (1) or not (0). I then recorded the highest correlation coefficient ($r_{highest}$). I repeated this process a large number of times (e.g., 1,000,000,000) and selected the final set of 40 items that had the lowest $r_{highest}$. For some participants, the self-descriptiveness and self-importance ratings were highly positively correlated. In such a case, I set a different criterion for that correlation. For instance, I computed $r_{highest}$ without considering $r_{self\text{-}descriptiveness/self\text{-}importance}$, and

selected a set of 40 items whose $r_{highest}$ was the lowest given that $r_{self\text{-}descriptiveness/self\text{-}importance}$ was <0.6. For examples of items, see Table 2.1.

Table 2.1 Examples of Items Used in the Experiments

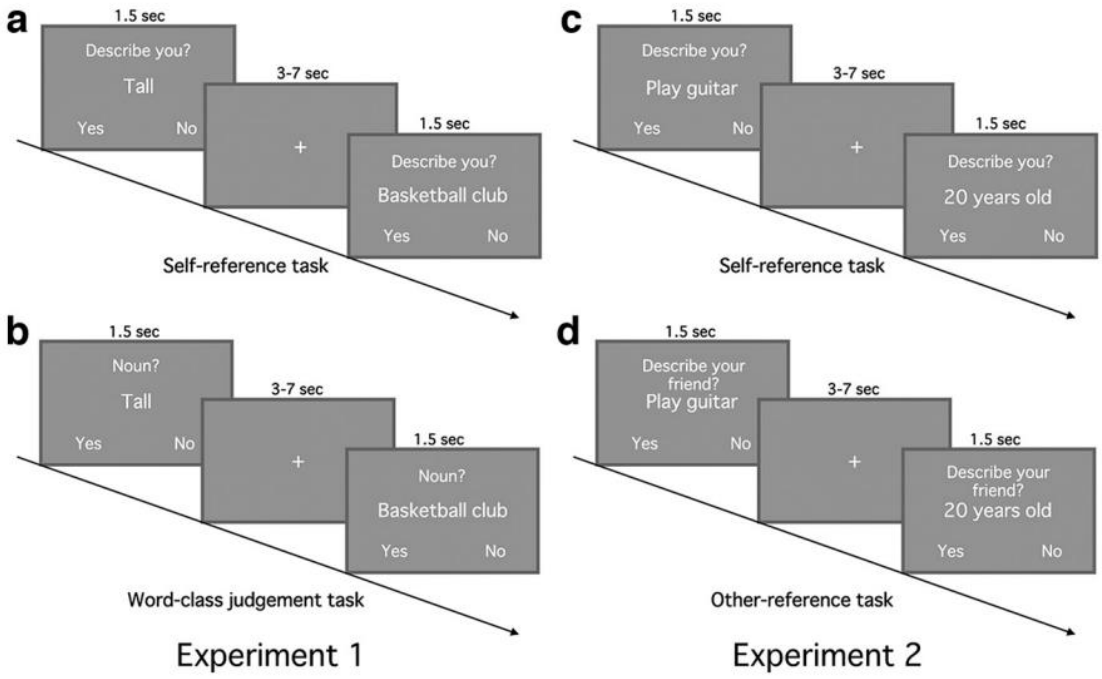
Physical	Social	Attributes	Other
Female	Art club	Talkative	Film
Bad eyesight	Flower arrangement club	Smart	Twitter
Tall	Department of Engineering	Hate prawns	Christmas
Hay fever	High School graduate	Open-minded	School trip
Born in Tokyo	Female	Compassionate	Rain
Sweaty	Softball team	Good singer	Piano
Brown hair	Japanese	Dog person	Earthquake
Sensitive skin	Basketball club	Play guitar	
Right-handed	Buddhist	Family-oriented	
20 years old	Kochi University of Technology student	Like to discuss ideas	

Note. I used 959 unique items across the two experiments. Items in the "Other" category were mainly prepared by an experimenter. The coding scheme is based on Cousins (1989).

2.3.2.4 fMRI Experiment

Before the fMRI scan, participants received instructions regarding MRI safety and tasks they would perform inside the fMRI scanner. During the fMRI session, participants performed two tasks: self-reference and word-class judgment (Figure 2.1*a,b*). I programmed both of them using Psychtoolbox (http://psychtoolbox.org/) with MATLAB software (version 2018a; http://www.mathworks.co.uk). Participants completed six runs of each task. Each run consisted of 40 trials, one item per trial. I presented the same set of 40 items in both tasks and in each run. I counterbalanced task (ABBAABBAABBA or BAABBAABBAAB), and randomized trial order within each run.

Figure 2.1 Experimental Tasks



Note. The self-reference task (a) and the word-class judgement task (b) in Experiment 1. The self-reference task (c) and other-reference task (d) in Experiment 2.

In the self-reference task, for each trial, participants viewed an item. On the screen, above the characteristic, they encountered the question “Describes you?”. For each trial, they could answer “yes” or “no” to indicate whether the characteristic described them or not (Figure 2.1a). For each trial in the word-class judgment task, participants viewed an item. On the screen, above the characteristic, they encountered the question “Noun?”. They could answer “yes” or “no” to indicate whether the characteristic was a noun or not (Figure 2.1b). For both tasks, each item was presented for 1.5 s, followed by intertrial interval (ITI; 3–7 s, mean = 5 s). Participants answered by pressing one of two buttons on a response box.

How vividly each item evoked a personal memory might differ between the second questionnaire and the fMRI task. Consequently, after the fMRI scan, I instructed participants to rate the same 40 items on autobiographical memory, namely, to what extent each item evoked an autobiographical memory when seeing it inside the fMRI scanner (1 = *it did not evoke any memory at all*, 7 = *it evoked very vivid memory*). Furthermore, to check for consistency of the self-descriptiveness judgment, I instructed participants to rate the 40 items again on self-

descriptiveness using the same response scale. Next, participants completed a demographic questionnaire.

2.3.2.5 fMRI Data Acquisition

I acquired images using a 3-T Trio A Tim MRI (Siemens) scanner with a 32-channel head coil. For functional imaging, I used T2*-weighted gradient-echo echo-planar imaging (EPI) sequences with the following parameters: time repetition (TR) = 2500 ms, echo time (TE) = 25 ms, flip angle (FA) = 90°, field of view (FOV) = 192 mm^2, matrix = 64 × 64. I acquired, in an interleaved order, 42 contiguous slices with a thickness of 3 mm. In addition, I acquired a T1-weighted structural image from each participant.

2.3.3 Statistical Analysis

2.3.3.1 Behavioural data analysis

Each participant rated each of the 40 items on self-descriptiveness, and they did so three times: (1) during the second online questionnaire; (2) during the fMRI scan; (3) after the fMRI scan. Although participants rated each item six times (across six fMRI runs) on a two-point scale (yes or no) during the fMRI scan, they rated each item once on a seven-point scale during the second questionnaire and post-fMRI rating task. Thus, for the self-descriptiveness rating data obtained during the fMRI scan, I computed a self-descriptiveness score for each item as a proportion of yes responses across six ratings of each item. I assumed that participants maintained a stable self-concept across a few weeks, and I tested this assumption by checking for consistency of their self-descriptiveness ratings obtained across the three times (or sessions).

2.3.3.2 fMRI Data Preprocessing

I conducted preprocessing and statistical analysis of the fMRI data using SPM12 (Welcome Department of Imaging Neuroscience), implemented in MATLAB (MathWorks). I discarded the first four volumes before preprocessing and data analyses to allow for T1

equilibration. I conducted preprocessing of the fMRI data with SPM 12's preproc_fmri.m script starting with realignment of all functional images to a common image. I spatially realigned all images within each run to the first volume of the run using seventh-degree B-spline interpolation, and I unwarped and corrected for motion artefacts. I segmented the T1-weighted structural image and normalized it into a common stereotactic space (MNI atlas). Subsequently, I applied the normalization parameters to the functional images and resampled them to 3 × 3 × 3 mm^3 isotropic voxels (i.e., original voxel size was retained) using seventh-degree B-spline interpolation. Following the normalization, I spatially smoothed the data [with a Gaussian kernel of 8-mm full-width at half-maximum (FWHM)] for the univariate analysis. To maintain fine grained activation patterns, I did not apply smoothing before the first-level data analysis for the RSA. I applied smoothing before the group analysis of the RSA outputs to account for individual variability in brain structure (with a Gaussian kernel of 4-mm FWHM).

2.3.3.3 fMRI Data Analysis: Univariate Analysis

I used three general linear models (GLMs) to analyse the fMRI data. In the first GLM, I compared the two conditions (self vs word), whereas I used the spmT images from the second GLM for the RSA. In the first GLM, I separately modelled 40 self-reference judgment trials and 40 word-class judgment trials using a box-car function convolved with the canonical hemodynamic response function.

In the second GLM, I investigate whether mPFC activities parametrically increase as a function of self-importance and/or self-descriptiveness ratings. As in the first GLM, I separately modelled 40 self-reference trials and 40 word-class judgment trials. In addition, I added to each of the self and word trial regressors the following seven parametric regressors: (1) self-descriptiveness; (2) self-importance; (3) valence; (4) familiarity; (5) autobiographical memory; (6) word-length; (7) whether each item was self-provided (1) or not (0). Given that SPM automatically performs orthogonalization for parametric regressors (Mumford et al., 2015), I also tried another GLM where the order of self-descriptiveness and self-importance parametric regressors were switched (self-importance as the first parametric regressor, and self-

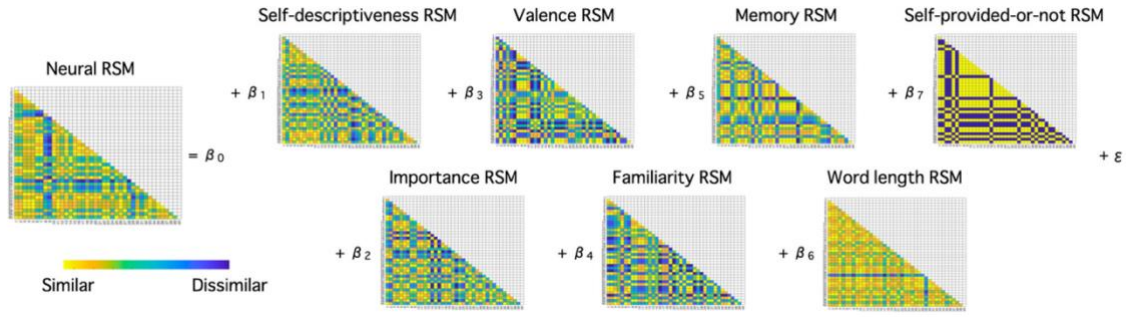
descriptiveness as the second parametric regressor), but the results were virtually the same. For the first two GLMs, I submitted the contrast images to a second level analysis. I set statistical threshold at $p < 0.005$ with cluster-$p < 0.05$ [familywise error (FEW) corrected] within the mPFC mask. Outside of the mask, I set up the statistical threshold at $p < 0.001$ (uncorrected for multiple comparisons) with a cluster threshold of $p < 0.05$ (FWE corrected).

In the third GLM, I modelled separately each of the 40 items for each task. I used a total of 80 spmT images from the third GLM in subsequent RSA. In all the GLMs, I included six head motion parameters and session effects as nuisance regressors.

2.3.4 Representational Similarity Analysis (RSA): Model Representational Similarity Analysis (RSM)

To test the effect of self-descriptiveness and self-importance on neural responses in the mPFC, controlling for other factors (i.e., valence, familiarity, autobiographical memory, word-length, whether items were self-provided or not), I used RSA with a searchlight approach. For each participant, I calculated a model RSM separately for the following seven dimensions: (1) self-descriptiveness; (2) self-importance; (3) valence; (4) familiarity; (5) autobiographical memory; (6) word-length; (7) whether each item was self-provided (1) or not (0). For self-descriptiveness, self-importance, valence, and familiarity, I used ratings from the second online questionnaire. For autobiographical memory, I used ratings from the postscan behavioural session. Each RSM was a 40 × 40 matrix (Figure 2.2), where each cell represented the similarity of the ratings between two items. For ratings completed on a seven-point scale, I calculated similarity as seven minus the absolute difference between two ratings. For the word length, I calculated similarity as the maximum number of characters in the 40 items minus the absolute difference in the number of characters between two items. Lastly, for whether items were self-provided or not, I coded similarity as 0 if an item was provided by the participant but the other item was not (or vice versa), whereas I coded similarity as 1 otherwise (i.e., both items were provided by the participants or both items were provided by the experimenter). I standardised the values with the respective mean and SD for each rating before regression analyses.

Figure 2.2 Neural and Model RSMs in the Multiple Regression Analysis in Experiment 1



Note. For each participant, I made seven RSMs based on participant ratings and item characteristics. The value in each cell indicates a similarity between a pair of items on a given dimension. I made neural a neural RSM for each searchlight. I conducted a multiple regression analysis for each searchlight with the seven model RSMs.

2.3.4.1 RSA: neural RSM

I extracted local patterns of neural activity from searchlights with a three-voxel radius, so that each searchlight consisted of a maximum of 123 voxels (and less on the edges of the brain). For each searchlight, I calculated voxel-by-voxel correlations between each pair of the 40 items, which resulted in a 40 × 40 neural RSM (Figure 2.2). Consequently, the correlation in neural activities between two items within a searchlight was represented by a cell in the respective neural RSM. I Fisher-Z transformed the correlation values before further analyses.

2.3.4.2 RSA: multiple regression analysis

In each searchlight, I conducted a multiple regression analysis, where the seven model RSMs were independent variables and the neural RSM was dependent variable (Figure 2.2). I repeated this analysis for every searchlight across the mPFC region of interest (ROI; see below for more information about the mPFC mask applied) and the whole brain, resulting in a β-map for each of the seven independent variables for each of the two tasks (i.e., a total of 14 β-maps for each participant).

2.3.4.3 RSA: Group Analysis

I entered the β-maps into a group-level analysis that I computed with permutation testing (i.e., one-sample *t* test with 5000 permutations) using the Statistical NonParametric Mapping (SnPM) toolbox for SPM (Nichols & Holmes, 2002). Among several brain regions previously implicated in self-processing (Denny et al., 2012; Murray et al., 2012; Qin and Northoff, 2011), I focused especially on the mPFC, as a lesion study indicated that this region is necessary for a stable and accurate self-concept (but is not critical for knowledge about another person; Marquine et al., 2016). Accordingly, I applied an mPFC mask to the analysis to limit the group-analysis to voxels within the a priori ROI. I created the mask with the WFU PickAtlas toolbox for SPM (Maldjian et al., 2003). The mPFC ROI mask included Frontal_Sup_medial_L, Frontal_Sup_medial_R, Frontal_Mid_Orb_L, Frontal_Mid_Orb_R, Rectus_L, Rectus_R, Cingulum_Ant_L, and Cingulum_Ant_R (dilation factor = 2), which I took from the Anatomical Automatic Labeling (AAL) masks implemented in the WFU pickatlas toolbox. I applied the same statistical threshold as the univariate analyses above [within the mPFC mask, $p < 0.005$ with cluster-$p < 0.05$ (FWE corrected), and outside the mask, $p < 0.001$ with cluster-$p < 0.05$ (FWE corrected)].

2.4 Materials and Methods - Experiment 2

2.4.1 Preregistration

I preregistered the hypotheses, sample size, data analytic plan, and exclusion criteria on the Open Science Framework (https://osf.io/agq3b). I followed the preregistration in all analyses reported below, unless otherwise noted.

2.4.2 Participants

As preregistered, final analyses included a sample of 35 undergraduate students (23 men, 12 women) at Kochi University of Technology, Japan, ranging from 18 to 23 years ($M = 19.66$, $SD = 1.64$). All of them were right-handed, and none had a history of psychiatric disorders. I

scanned eight additional participants but excluded them from the final analyses, because they did not meet the preregistered inclusion criteria. In particular, I excluded one participant because of a brain anomaly, and seven participants because the reliability of either their self-reference or other-reference rating was low. In regard to these seven participants, I calculated correlations between their responses in the self-reference task during fMRI scanning and their self-descriptiveness rating in the second questionnaire, as well as the correlation between their other-reference task responses during fMRI scanning and their friend-descriptiveness rating in the second questionnaire. I considered the correlation value low, if it was <0.5. All participants ticked a box to indicate their consent before the online questionnaires, and they consented in writing before the fMRI experiment. The experiment was approved by the Kochi University of Technology ethics committee. They were remunerated with 2000 Japanese yen.

2.4.3 Power Analysis

I conducted a power analysis using the Bootstrap procedure. First, I randomly sampled 35 participants from the 28 participants of Experiment 1, with replacement. For each randomly-selected sample, I conducted a group analysis (one-sample *t* test). Given that in Experiment 1 I found significant activations within the mPFC, I applied the same mPFC mask created via the WFU PickAtlas toolbox for SPM (Maldjian et al., 2003). I applied a voxel-wise threshold of $p < 0.005$ (uncorrected) and cluster-$p < 0.05$ (FWE corrected) to assess significance. I repeated these steps 2000 times, and counted the number of times I found significant activations within the mPFC mask. The result indicated that $N = 35$ would achieve power of 91.85%.

2.4.4 Experimental Procedure

The procedure was similar to Experiment 1's, consisting of three parts (two online questionnaires, fMRI experiment) on three separate days. The only alteration involved the control task. To examine whether the Experiment 1 results were specific to the self, or whether the mPFC also encodes important information for a friend's identity, I used an other-reference task as control (Figure 2.1*c,d*). The first and second online questionnaires were separated by an

average of 11.0 d (SD = 6.68). The second online questionnaire and fMRI experiment were separated by an average of 14.03 d (SD = 8.34).

2.4.4.1 Online Questionnaires

As in Experiment 1, in the first online questionnaire, participants provided at least 30 characteristics by responding to the prompt "I ____." Similarly, participants provided the name of a close friend and at least 30 characteristics they believed to be descriptive of or important for that friend. They did so by responding to the prompt "My friend _____."

In the second online questionnaire, participants rated 80 items, some of which were made available by participants during the first online questionnaire. In particular, they rated each item on the following seven dimensions: (1) self-descriptiveness (1 = *not at all descriptive*, 7 = *very descriptive*); (2) self-importance (1 = *not at all important*, 7 = *very important*); (3) friend self-descriptiveness (1 = *not at all descriptive*, 7 = *very descriptive*); (4) importance to friend's identity (1 = *not at all important*, 7 = *very important*); (5) valence (1 = *very negative*, 7 = *very positive*), (6) familiarity (1 = *not at all familiar*, 7 = *very familiar*); (7) autobiographical memory (1 = *it did not evoke any memory at all*, 7 = *it evoked very vivid memory*). I selected a stimulus set of 40 items as in Experiment 1 (for item examples, see Table 2.1).

2.4.4.2 fMRI Experiment

During the fMRI session, participants conducted the self-reference and other-reference tasks (Figure 2.1c,d). I used the same set of 40 items for both tasks.

Just like in Experiment 1, during the self-reference task, for each trial, the participants viewed one of the 40 items. On the screen, above the characteristic, they saw the question "Describes you?" For each trial, they answered "yes" or "no" to indicate whether the characteristic described them (Figure 2.1*c*). For each trial in the other-reference task, participants similarly viewed an item on the screen. Above the item, they saw the question "Describes your friend?" and answered "yes" or "no" to indicate whether the characteristic described their friend, the same close friend they mentioned during the first online

questionnaire (Figure 2.1*d*). For both tasks, each item was presented for 1.5 s, followed by intertrial interval (ITI; 3–7 s, mean = 5 s). Participants indicated their answers by pressing one of two buttons on a response box.

2.4.4.3 Postscan Behavioural Session

After the scan, participants rated the previously presented words on autobiographical memory again (1 = *it did not evoke any memory at all*, 7 = *it evoked very vivid memory*). Next, they completed a demographic questionnaire.

2.4.4.4 fMRI Data Acquisition

I acquired images using a Siemens 3.0 T Verio MRI scanner with a 64-channel phased array head coil. For functional imaging, I used T2*-weighted gradient-echo echo-planar imaging (EPI) sequences with the following parameters: time repetition (TR) = 2500 ms, echo time (TE) = 25 ms, flip angle (FA) = 90°, field of view (FOV) = 192 mm^2, matrix = 64 × 64. I acquired 42 contiguous slices with a thickness of 3 mm, in an interleaved order. Moreover, I acquired from each participant a high resolution anatomic T1-weighted image (1-mm isotropic resolution).

2.4.5 Statistical Analysis

2.4.5.1 fMRI Data Processing

I conducted preprocessing of fMRI data as in Experiment 1. The preprocessing described in our preregistration stated that I would use an EPI-template when normalizing fMRI data to the standard MNI space. Although, based on visual inspection of normalized images, there was no issue with this method when analysing the fMRI data from Experiment 1, I noticed that fMRI images normalized with this method were consistently smaller in the anterior-to-posterior and left-to-right dimensions (possibly because of the difference in head-coil between the two experiments; 32 channels in Experiment 1 vs 64 channels in Experiment 2 (for a similar case, see Smith et al., 2018). Accordingly, I decided to use a T1-template when normalizing the fMRI data as implemented in the SPM 12's preproc_fmri.m script. For the sake of consistency, I re-

analysed fMRI data in Experiment 1 with this new preprocessing steps, as reported above. In Experiment 1, I report the re-analysed data (note that the two preprocessing steps generated virtually identical results).

2.4.5.2 Univariate fMRI Analysis

Similar to Experiment 1, I used three GLMs. In the first GLM, I intended to compare the two conditions (self vs other). In the second GLM (not preregistered), I intended to test whether mPFC activities increase parametrically as a function of self-descriptiveness and self-importance ratings. Finally, I used the spmT images from the third GLM for the RSA.

In the first GLM, I separately modelled 40 self-reference judgment trials and 40 other-reference trials using a box-car function convolved with the canonical hemodynamic response function. In the second GLM, as in the first one, I separately modelled 40 self-reference trials and 40 other-reference judgment trials. In addition, I added the following nine parametric regressors to each of the self-reference and other-reference trial regressors: (1) self-descriptiveness; (2) self-importance; (3) friend-descriptiveness; (4) friend-importance; (5) valence; (6) familiarity; (7) autobiographical memory; (8) word-length; (9) whether the item was self-provided or not.

For the first two GLMs, I submitted the contrast images to a second level analysis. As preregistered, I employed the same mPFC mask as in Experiment 1, and within the mPFC mask I set the statistical threshold at $p < 0.005$ (uncorrected for multiple comparisons) with a cluster threshold of $p < 0.05$ (FWE corrected). Outside of the mask, I set the statistical threshold at $p < 0.001$ (uncorrected for multiple comparisons) with a cluster threshold of $p < 0.05$ (FWE corrected).

In the third GLM, I modelled separately each of the 40 items for each task. In all three GLMs, I included six head motion parameters and session effects as nuisance regressors.

2.4.5.3 Model RSMs

I conducted searchlight RSA as in Experiment 1. However, in addition to testing the effect of self-descriptiveness and self-importance, I tested the effect of friend-descriptiveness and friend-importance, on neural representations. So, for each participant, I calculated a model RSM separately for each of the following nine dimensions: (1) self-descriptiveness; (2) self-importance; (3) friend-descriptiveness; (4) friend-importance; (5) valence; (6) familiarity; (7) autobiographical memory; (8) word-length; (9) whether the item was self-provided or not. For self-descriptiveness, self-importance, friend-descriptiveness, friend-importance, valence, and familiarity, I used the ratings from the second questionnaire. For autobiographical memory, I used the ratings from the postscan behavioural session.

2.4.5.4 Neural RSM

I created a neural RSM for each searchlight as in Experiment 1.

2.4.5.5 RSA: Multiple Regression Analysis

In each searchlight, I conducted a multiple regression analysis where the nine model RSMs were independent variables and the neural RSM was the dependent variable. I repeated the analysis for every searchlight across the brain, resulting in a β-map for each of the nine independent variables and each of the two tasks [a total of 18 (2 × 9) β-maps for each participant]. Although not preregistered, I attempted another RSA by adding a model RSM based on participants' average RT for each item (a total of 10 model RSMs), and this additional RSA produced results virtually identical to those reported below.

2.4.5.6 RSA: Group Analysis

I conducted the second-level group analysis as in Experiment 1 (i.e., using SnPM). I applied the same statistical threshold as the univariate analyses above [within the mPFC mask, $p < 0.005$ with cluster-$p < 0.05$ (FWE corrected), and outside the mask, $p < 0.001$ with cluster-$p < 0.05$ (FWE corrected)].

2.4.5.7 Classifier-Based MVPA (not Preregistered)

I also conducted a classifier-based MVPA analysis that directly compares the effects of self-importance and friend-importance on mPFC activation. I did so in search for evidence that a neural code for information importance is unique to the self. Specifically, I tested whether, during the other-reference task, the mPFC activation patterns evoked by items high (also middle or low) in self-importance task are distinct from activation patterns evoked by items high (also middle or low) in friend-importance.

First, I conducted another GLM analysis where each item was classified into one of the three categories depending on level of self- (and friend-)importance: high, middle, low. Given that the distribution of ratings was different across participants (e.g., with some frequently providing ratings of 6–7, and others frequently providing ratings of 1–2), I used different criteria for different participants when classifying each item into the three categories so that the three categories included roughly an equal number of items. Of note, within each participant, I used the same criterion for the self and other conditions. Thus, in this GLM, when modelling the fMRI data from the self-reference task, I classified 40 items into three categories based on self-importance ratings: (1) self-importance-high; (2) self-importance-middle; (3) self-importance-low. I modelled separately items in each of the three categories. Similarly, for the other-reference task fMRI data, I classified items into three categories in the same way based on the friend-importance ratings (high, middle, or low), and I modelled separately items in each of the three categories. I included six head motion parameters and session effects as nuisance regressors. I then computed a spmT map for each category per fMRI run resulting in 2 (tasks; self vs other) × 3 (level of importance) × 6 (runs) spmT images per participants, which I used in the subsequent MVPA.

To define independently a self-importance related mPFC ROI, I used a leave-one-participant-out cross-validation procedure (Esterman et al., 2010). I re-ran the second-level group analysis (the searchlight RSA group analysis described above) 34 times with a different single participant left out in each. I used each second-level analysis to determine an mPFC ROI

for each left-out participant. For each participant, I extracted data from a three-voxel radius sphere surrounding the peak voxel within the mPFC most strongly associated with self-importance ratings. To ascertain that each participant's ROI was roughly from the same anatomic subregion within the mPFC, I searched a peak voxel for each participant within a 30-mm sphere surrounding the peak voxel identified by the group analysis with all 35 participants.

I used a linear support vector machine, which I conducted using Matlab in combination with LIBSVM (https://www.csie.ntu.edu.tw/~cjlin/libsvm/; Wake and Izuma, 2017) with a cost parameter of c = 1 (default). I paired each of the self run and friend run in the order of acquisition, and evaluated classification performances with a leave-one-pair-out cross-validation procedure. Thus, using the spmT images from the five runs of each task, I trained a classifier that discriminates activation patterns between self-importance-high versus friend-importance-high items. Then, using the spmT images from the left-out run of each task, I tested whether the classifier could discriminate between self-importance-high versus friend-importance-high items. I repeated the procedure six times so that each run-pair served as the testing set once. I averaged six classification accuracy values for each participant. I conducted the same analysis to test whether activation patterns are distinct between items low in self-importance versus items low in friend-importance (also, items middle in self-importance vs middle in friend-importance). I assessed statistical significance with permutation tests where I performed classifications with scuffled labels 1000 times to obtain a null distribution; *p*-values were set at 0.05 (one-tailed) and Bonferroni-corrected for three comparisons.

2.4.6 Data Availability

Unthresholded group-level statistical maps and the mPFC mask image are available on NeuroVault (https://neurovault.org/collections/13069/).

2.5 Results – Experiment 1

2.5.1 Behavioural Results

Participants rated each of the 40 items on self-descriptiveness three times: (1) during the second online questionnaire; (2) during the fMRI scan; (3) after the fMRI scan. Their responses were highly consistent across the three sessions (average within-individual correlation = 0.74; Table 2.2). This finding supports our assumption that participants' self-concept was stable over the weeks of testing.

Table 2.2 Average Within-Person Correlations (SD) Across the Three Self-Descriptiveness Ratings in Experiment 1

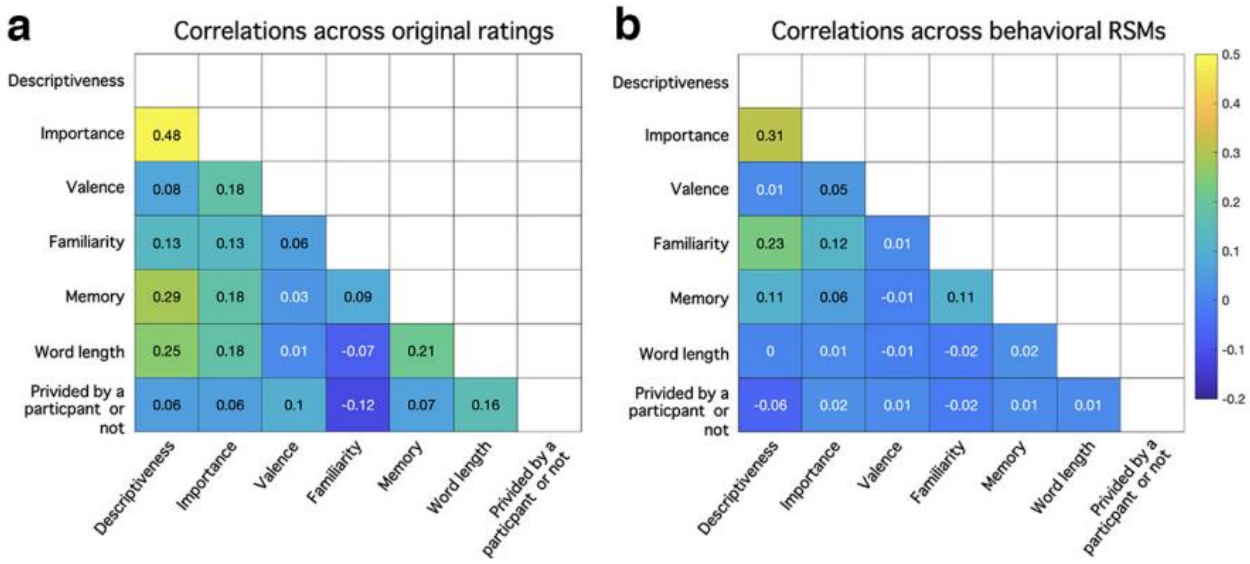
	Second questionnaire	fMRI task	Post-fMRI rating
Second questionnaire	-		
fMRI task	0.83 (0.09)***	-	
Post-fMRI rating	0.72 (0.14)***	0.67 (0.12)***	-

Note. Each participant rated each of the 40 items on self-descriptiveness three times; 1) during the second online questionnaire, 2) during the fMRI scan, and 3) after the fMRI scan. ***$p < 0.001$ (corrected for multiple comparisons) based on one sample t-test (one-tailed; correlation coefficients were Fisher-z transformed before the t-tests).

I present average correlations between the behavioural ratings (self-descriptiveness, self-importance, valence, familiarity, autobiographical memory, word-length, whether items were self-provided) from the second questionnaire in Figure 2.3a (note that memory ratings were from the second memory rating task after the fMRI scan). Similarly, I present average correlations across the seven model RSMs in Figure 2.3b. The correlation between the self-descriptiveness and self-importance model RSMs was the highest [average $r = 0.31$ ($SD = 0.26$)]. I also calculated and checked the variance inflation factors (VIFs) for the seven independent variables within each participant. VIF provides an index of the degree to which the variance of a coefficient is increased because of collinearity, with values of above 10 often considered problematic. Across a total of 196 (7 variables × 28 participants) VIFs, 193 of them were below 2,

and the maximum VIF was 3.03, indicating reasonable ability to draw inferences on the unique variance explained by each variable in all participants.

Figure 2.3 Average Correlations Across Ratings and Model RSM



Note. Average correlations across seven ratings (a) and seven model RSMs (b) in Experiment 1.

2.5.2 fMRI Results

2.5.2.1 Univariate Analysis

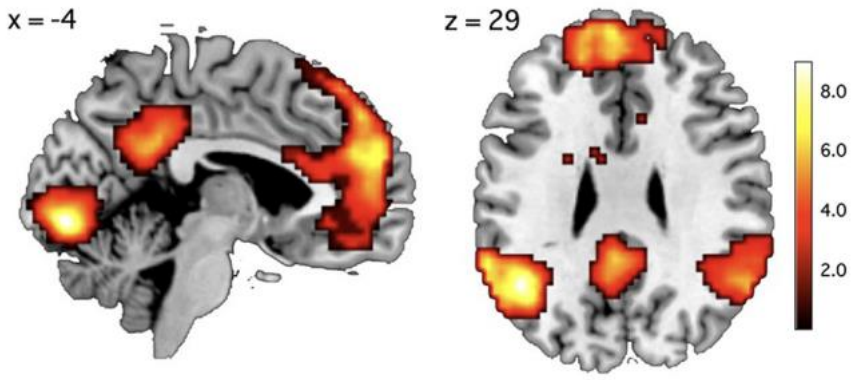
Successfully replicating the previous experiments (Qin & Northoff, 2011; Denny et al., 2012; Murray et al., 2012), I found that the self-reference versus word-class judgment contrast significantly activated the mPFC and PCC (Figure 2.4). Other activated regions included left and right temporoparietal junction (TPJ), left superior temporal sulcus (STS), and lingual gyrus (Figure 2.4; Table 2.3). The opposite contrast (word versus self) activated the left inferior frontal gyrus (IFG), which is known to play a major role in language processing (Ferstl et al., 2008; Table 2.3)

Table 2.3 Brain Regions Showing Significant Activations During the Self-Reference Task and the Word-Class Judgement Task in Experiment 1

Contrast	Location	MNI coordinates			Z	Cluster size (voxels)
		x	y	z		
Self > Word	dmPFC	-9	41	53	5.62	1,605
	amPFC	-3	50	23	5.42	
	dACC	6	20	20	4.19	
	MFG	-33	20	38	4.68	171
	left STS	-60	-22	-10	5.55	736
	PCC	-3	-46	29	4.82	370
	left TPJ	-45	-58	29	6.20	580
	right TPJ	57	-58	29	5.83	339
	Lingual gyrus	-3	-85	-4	6.37	722
Word > Self	left IFG	-48	32	20	4.69	519

Note. The statistical threshold was set at $p < 0.005$ (uncorrected for multiple comparisons) with a cluster threshold $p < 0.05$ (FWE corrected). dmPFC; dorsomedial prefrontal cortex, amPFC; anterior medial prefrontal cortex, dACC; dorsal anterior cingulate cortex, MFG; middle frontal gyrus, STS; superior temporal sulcus, PCC; posterior cingulate cortex, TPJ; temporoparietal junction, IFG; inferior frontal gyrus. Voxel size = 3 × 3 × 3 mm.

Figure 2.4 Group Activation Maps for Self Versus Word Contrast in Experiment 1



Note. For a display purpose, I set the voxel-wise threshold at $p < 0.005$ (uncorrected) and set the cluster size threshold at $p < 0.05$(FWE corrected). For all activated areas, see Table 2.3.

2.5.2.2 Parametric modulation analysis

Next, I tested whether mPFC activities parametrically increase as a function of self-descriptiveness and/or self-importance. Contrary to previous experiments (D'Argembeau et al., 2012; Elder et al., 2022; Koski et al., 2020; Macrae et al., 2004; Moran et al., 2006; Koski et al., 2020), neither self-descriptiveness nor self-importance ratings were significantly associated

with mPFC activations during the self-reference task. Other experiments with a similar event-related design did not report results relevant to this association (Heatherton et al., 2006; Kelley et al., 2002; Ochsner et al., 2005; Yaoi et al., 2015), although they could have done so. Furthermore, the results of a recent experiment with direct neural recordings from human participants using electrocorticography (ECoG; Tan et al., 2022) suggest that the linear relationship between self-descriptiveness and mPFC activations is, if anything, small. Both self-descriptiveness and self-importance ratings were unrelated to mPFC activities during the word task.

2.5.2.3 Searchlight RSA Result

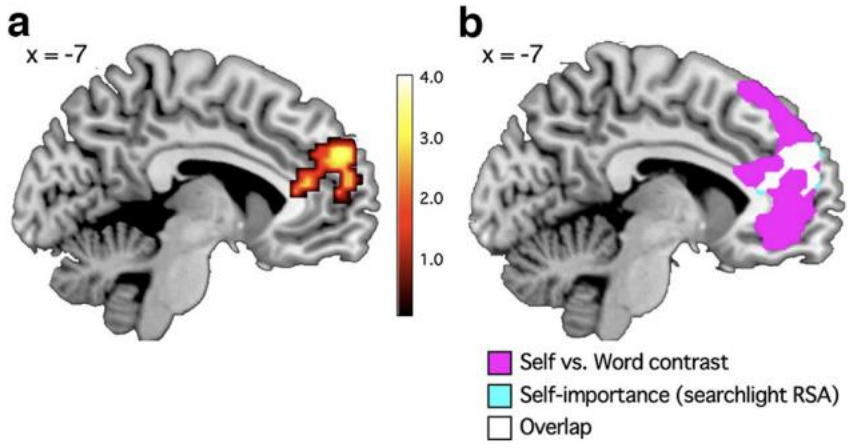
I conducted searchlight RSA within the mPFC ROI to test whether self-descriptiveness or self-importance had an effect on the local patterns of activation within the mPFC. I found that different levels of self-importance were represented by different patterns of activation within the mPFC (medial superior frontal gyrus; $x = -9$, $y = 53$, $z = 29$, 306 voxels) during the self-reference task (Figure 2.5; Table 2.4). However, self-descriptiveness was not significantly associated with activation patterns within the mPFC. Likewise, the remaining five variables were not significantly associated with mPFC activations. The mPFC region associated with self-importance (Figure 2.5a) largely overlapped with the mPFC region activated by the self versus word contrast (Figure 2.4). Out of the 306 voxels whose activities were significantly associated with self-importance, 194 voxels (63.3%) were included in the area significantly activated by the self-reference task compared with the word task (Figure 2.5b).

Table 2.4 mPFC Regions from Searchlight RSA Showing Significant Association with Self-Importance Ratings During the Self-Reference Task in Experiment 1

Location	MNI coordinates			Z	Cluster size (voxels)
	x	y	z		
dmPFC	-9	53	29	3.84	306
dACC	12	38	23	3.47	
amPFC	6	56	14	2.89	

Note. The statistical threshold was set at $p < 0.005$ (uncorrected for multiple comparisons) with a cluster threshold $p < 0.05$ (FWE corrected). Voxel size = 3 × 3 × 3 mm.

Figure 2.5 Searchlight RSA Results in Experiment 1



Note. (a) Self-importance was significantly associated with activation patterns within the mPFC during the self-reference task (see also Table 2.4). $p < 0.005$ (uncorrected) and cluster-$p < 0.05$ (FWE corrected). (b) The mPFC areas significantly associated with self-importance (cyan) largely overlapped with the areas activated by the self-reference task compared to the word-class judgement task (magenta; Figure 2.4).

I repeated the same searchlight RSA using the data from the word-class judgment task, and found no significant results. The null effects indicate that the representation of self-importance within the mPFC is task dependent. Self-importance is represented within the mPFC only when performing a task that requires thinking about the self.

Outside of the mPFC ROI, different levels of word-length were represented by different patterns of activation within the visual cortex (lingual-gyrus) for both the word-class judgment task and the self-reference task, indicating that visually similar stimuli evoke similar activation patterns in the visual cortex regardless of task. No other significant results emerged.

Taken together, I obtained initial evidence that the mPFC represents self-importance information. However, it is possible that the mPFC represents importance not specific to self-identity, but relevant to another person's identity as well; that is, the mPFC may not be specific to the self, but instead process person information in general. I addressed this possibility in Experiment 2.

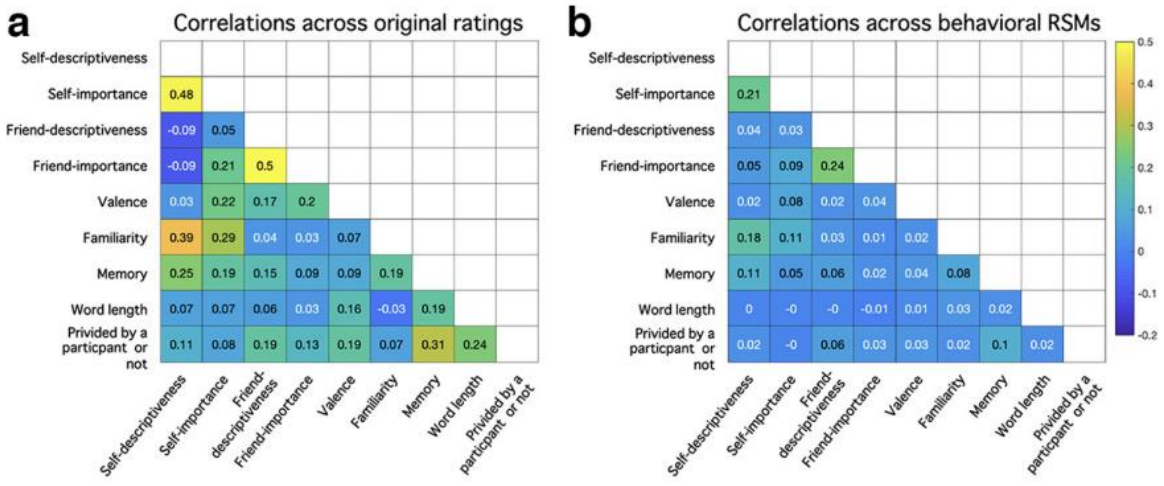
2.6 Results – Experiment 2

2.6.1 Behavioural Results

Participants rated each of the 40 items on self-descriptiveness and friend-descriptiveness twice: (1) during the second online questionnaire; (2) during the fMRI scan. Their responses were highly consistent across the two sessions. Average within-individual correlation for self-descriptiveness ratings was 0.78 (*SD* = 0.09), and average within-subject correlation for friend-descriptiveness was 0.74 (*SD* = 0.09).

I present average correlations between the behavioural ratings (self-descriptiveness, friend's self-descriptiveness, self-importance, friend's self-importance, valence, familiarity, autobiographical memory, word-length, whether items self-provided) from the second questionnaire in Figure 2.6. I checked the VIFs for the nine independent variables within each participant. Results showed that all 315 (9 variables × 35 participants) VIFs were below 2, indicating reasonable ability to make inferences on the unique variance explained by each variable in all participants.

Figure 2.6 Average Correlations Across Ratings and Model RSMs



Note. Average correlations across nine ratings (a) and nine model RSMs (b) in Experiment 2.

2.6.2 fMRI Results

2.6.2.1 Univariate Analysis

The self-reference versus other reference contrast did not reveal significant activation within the mPFC or across the whole brain. Although this result is in contrast to our preregistered hypothesis, previous experiments have generated mixed findings regarding the difference between the self and other conditions, and our finding is consistent with experiments that reported no difference (Benoit et al., 2010; Ochsner et al., 2005; Schmitz et al., 2004; Tan et al., 2022; Vanderwal et al., 2008). The opposite contrast also did not reveal significant activation in any region.

2.6.2.2 Parametric Modulation Analysis

I investigated whether mPFC activities parametrically increase as a function of self-descriptiveness and/or self-importance. However, as in Experiment 1, neither self-descriptiveness nor self-importance ratings were significantly associated with mPFC activations during the self-reference task. Similarly, neither friend-descriptiveness nor friend-importance were significantly associated with mPFC activations during the other-reference task.

2.6.2.3 Searchlight RSA within the mPFC ROI

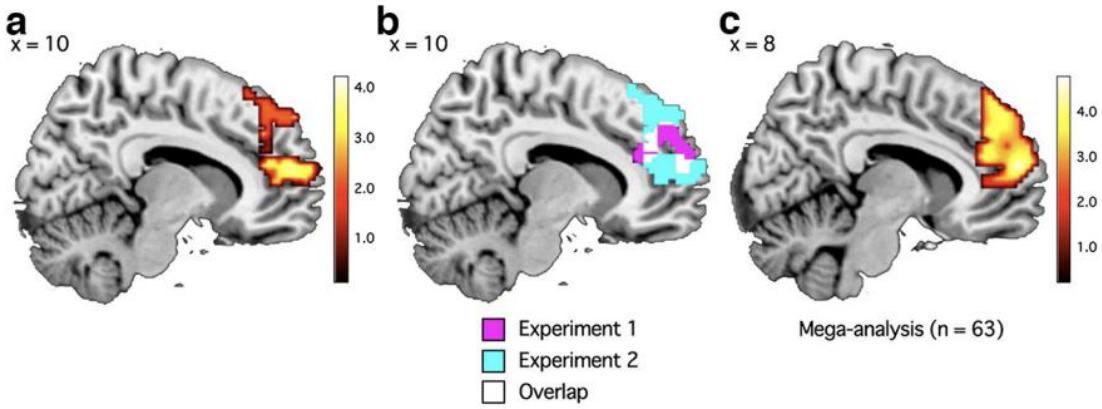
Based on our preregistered hypothesis that self-importance is encoded in the mPFC, I first limited the search area to within the mPFC by applying the anatomic mPFC mask. I conducted searchlight RSA to test whether self-importance information is represented in areas within the mPFC during self-reference task. As hypothesised, self-importance was reliably signalled in the mPFC during the self-reference task (x = 3, y = 41, z = 50; 280 voxels; Figure 2.7a; Table 2.5). This mPFC cluster overlapped with the mPFC cluster related to self-importance in Experiment 1, although the overlap was relatively small (a total of 25 voxel; Jaccard index = 0.046; Figure 2.7b).

Table 2.5 mPFC Regions from Searchlight RSA Showing Significant Association with Self-Importance During the Self-Reference Task in Experiment 2

Location	MNI coordinates			Z	Cluster size (voxels)
	x	y	z		
dmPFC	3	41	50	3.65	280
amPFC	9	53	14	3.43	
dACC	9	41	17	3.02	

Note. The statistical threshold was set at $p < 0.005$ (uncorrected for multiple comparisons) with a cluster threshold $p < 0.05$ (FWE corrected). Voxel size = 3 × 3 × 3 mm.

Figure 2.7 Searchlight RSA Results in Experiment 2



Note. (a), activation overlaps between Experiments 1 and 2 (b), and mega-analysis results (c). (a) Self-importance was significantly associated with activation patterns within the mPFC in Experiment 2 (see also Table 2.5). $p < 0.005$ (uncorrected) and cluster-$p < 0.05$ (FWE corrected). (b) mPFC areas associated with self-importance in Experiment 2 (magenta) overlap with areas associated with self-importance in Experiment 1 (cyan). (c) Mega-analysis results ($n = 63$) showing an mPFC cluster that is significantly associated with self-importance (see also Table 2.7). $p < 0.005$ (uncorrected) and cluster-$p < 0.05$ (FWE corrected).

In contrast, self-descriptiveness was not encoded in the mPFC during the self-reference task. Furthermore, neither self-importance nor self-descriptiveness were encoded in the mPFC during the other-reference task. These results replicate those of Experiment 1. Processing of information about how important each stimulus is to the self in the mPFC is task dependent, and its neural representations emerge only when performing a task that requires thinking about the self.

In contrast, both friend-descriptiveness and friend-importance were not significantly associated with mPFC activation during the self-reference and other-reference tasks. Likewise, the remaining five variables were not significantly related to mPFC activation during either task.

2.6.2.4 Whole-brain searchlight RSA

I performed searchlight RSA throughout the whole-brain. Consistent with Experiment 1, I found that different levels of word-length were represented by different patterns of activation within the visual cortex (lingual-gyrus) for both the self-reference and other-reference tasks. I also found that, during the self-reference task, familiarity ratings were related to activation patterns in left middle frontal gyrus (MFG; $x = -21$, $y = 4$, $z = 47$, 648 voxels) and in left inferior frontal gyrus (IFG; $x = -45$, $y = 11$, $z = 14$, 302 voxels). No other significant effects emerged.

2.6.3 Exploratory Analysis Directly Comparing Effects of Self-Importance and Friend-Importance

Although classification performance for items high in importance was not significant, the classifier-based MVPA successfully discriminated activation patterns between self-importance-middle versus friend-importance-middle and between self-importance-low versus friend-importance-low (Table 2.6), indicating that neural codes for information importance are largely unique to the self. Classification performance was not significant for items high in importance; however, our additional analysis found that the data were noisier for items high in importance (i.e., activation patterns evoked by the same item were less consistent across six runs of the same task; Figure 2.8). It also found that the self-reference task data were noisier compared with those of the other-reference task. Thus, mPFC activation patterns evoked by items high in self-importance during the self-reference task were least consistent (i.e., noisiest) across six runs. This might be because performing the self-reference (vs other-reference) task and items high (vs low) in importance evokes other cognitive/affective processes (e.g., autobiographical memory, positive affect) that can influence mPFC activation (Bartra et al., 2013; Martinelli et al., 2013), and these unrelated processes might have impacted on mPFC activation patterns

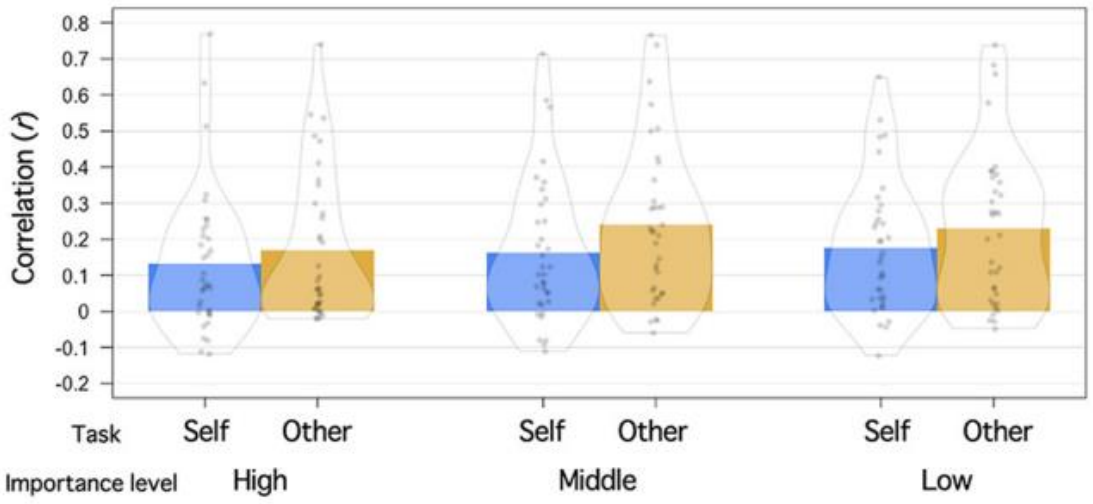
differently in each run. Regardless, the elevated level of noise observed in items high in self-importance explains, at least partially, the nonsignificance classification performance for items high in self- versus friend-importance.

Table 2.6 Classifier-based MVPA Results

Comparison	Classification Performance	P_{perm} (uncorrected)
High importance	50.48%	0.40
Middle importance	62.62%	< 0.001**
Low importance	55.71%	0.003*

Note. * $p < 0.05$ and ** $p < 0.01$ (Bonferroni corrected). Significant classification performance means that activation patterns were distinct in the self versus other conditions for a given importance level.

Figure 2.8 Average Within-Condition Correlations



Note. Average within-condition correlations for each of the six conditions (2 [task; self versus other] × 3 [importance-level; high, middle, or low]). Each participants completed six runs of each task, and within-condition correlations were computed for all possible run-pairs (a total of 15) which was averaged within each participant. A 2 × 3 repeated-measures Analysis of Variance (ANOVA) revealed significant main effects of both task ($F_{(1, 34)} = 6.71$, $p = 0.014$) and importance-level ($F_{(2, 68)} = 3.67$, $p = 0.0307$), whereas a task × importance-level interaction was not significant ($F_{(2, 68)} = 0.5262$ $p = 0.539$). Note that the correlation coefficients were Fisher-z transformed before conducting the ANOVA.

2.6.4 Mega-Analysis (not preregistered)

Given that the self condition was common across Experiments 1 and 2, I combined the self-task data from both experiments and ran a mega-analysis that included 63 participants. Self-importance information was reliably encoded in a large cluster in the mPFC (x = 15, y = 38, z = 26, 942 voxels; Figure 2.7c; Table 2.7).

Table 2.7 mPFC Regions from Searchlight RSA (Mega-Analysis; n = 63) Showing Significant Association with Self-Importance During the Self-Reference Task

Location	MNI coordinates			Z	Cluster size (voxels)
	x	y	z		
dACC	15	38	26	4.42	942
amPFC	9	56	14	4.34	
dmPFC	6	47	38	3.82	

Note. I set up the statistical threshold at $p < 0.005$ (uncorrected for multiple comparisons) with a cluster threshold $p < 0.05$ (FWE corrected). Voxel size = 3 × 3 × 3 mm.

When I did not apply the anatomic mPFC mask, the above mPFC cluster extended laterally to right IFG (x = 36, y = 38, z = 5) and left IFG (x = −33, y = 29, z = 29) consisting of 2087 voxels. I observed no other significant cluster for self-importance. Familiarity information was represented in right superior frontal gyrus (SFG; x = 27, y = 26, z = 53; 303 voxels), and autobiographical memory information was represented in right lingual gyrus (x = 27, y = −67, z = −1; 440 voxels). Furthermore, unsurprisingly, word length was strongly associated with activation patterns in the visual cortex (right: x = 12, y = −85, z = 2, and left: x = −9, y = −91, z = 2; a total of 2413 voxels). I observed no significant result for self-descriptiveness and valence.

2.7 Discussion

Previous research has linked the self-reference task to neural activation in the mPFC (Wagner et al., 2012), but it is unclear which information about the self is represented in that brain region. Across two self-reference experiments, controlling for potential confounds, I consistently demonstrated that the mPFC represents how important attributes are to one's self-identity. The results suggest that the self-concept is represented in the mPFC and

conceptualized in terms of self-importance, not self-descriptiveness. Furthermore, in both experiments, the parametric modulation analysis found no significant activation in the mPFC. Thus, these results indicate that self-importance information systematically alters activation patterns, but does not affect overall activation magnitude in the mPFC. In Experiment 2, I did not observe the relationship between mPFC neural responses and importance in the other (best friend) condition, and mPFC activation patterns associated with each of three levels of importance were generally distinct between self and best friend, suggesting that the mPFC represents information about the importance of information specifically to the self. Taken together, our research improves understanding about how and where the self is represented in the brain.

Although I found an association between self-importance and mPFC activation patterns across two experiments, the self-importance sensitive mPFC areas did not overlap widely (Figure 2.7*b*). Given that locations of peak self-related activations reported in previous neuroimaging experiments vary greatly along the z-coordinates (e.g., from $z = -10$ to $z = 70$; see also Denny et al., 2012), there might be considerable individual differences in functional dissociations within the mPFC. Thus, only focusing on a group average may prevent researchers from fully understanding the role of the mPFC in the self-concept (and the mPFC functions more generally). Follow-up research should consider and address this possibility.

Although neuroimaging research has documented the involvement of the default mode network in the self-reference task (Qin & Northoff, 2011; Wen et al., 2020), the current results indicate that only the mPFC is associated with self-importance. This finding is consistent with a previous lesion study, which illustrated the mPFC's crucial role in accurate and reliable trait knowledge of the self (Marquine et al., 2016). A patient (J.S., 74-year-old white male) had extensive damage to the medial prefrontal areas including orbitofrontal cortex and anterior cingulate gyrus. He and control participants completed a self-reference task on two occasions using the same trait adjectives. A male nurse who had known Patient J.S. for five years also rated patient J.S. on the same traits. Patient J.S.'s ratings were less consistent across two

sessions and less consistent with ratings done by the nurse compared with the control group. On the other hand, when patient J.S. was asked to rate the nurse, his ratings were consistent across two sessions and consistent with ratings done by the nurse himself, indicating that trait knowledge of another person was preserved. In similar experiments with various patients (e.g., autism, ADHD, Alzheimer's disease), trait self-knowledge was remarkably resistant to neural and cognitive damage (Klein and Lax, 2010). Thus, to the best of our knowledge, damage to the mPFC is the only case where trait knowledge of self is impaired, which is in a sharp contrast to other nonself-related knowledge that is impaired after damage to parietal, temporal, or frontal areas (Damasio et al., 2004; Gainotti, 2000; Neininger & Pulvermüller, 2003).

I demonstrated across two experiments that self-importance is represented in the mPFC only during the self-reference task, whereas stimulus' perceptual properties (i.e., word length) are represented in the visual cortex regardless of the task involved. This task-specific neural representation is consistent with RSA experiments on object representations (Bracci et al., 2017), which showed that information relevant to a given task is represented in prefrontal and parietal areas only while performing the task, whereas occipitotemporal areas mainly represent stimulus' perceptual properties (e.g., object shape) regardless of task involved. However, during the self-reference task, participants judged whether each personality trait describes them or not; this task does not explicitly require judging how important each trait is to one's identity. Hence, our results suggest an interesting possibility: individuals may actively use self-importance information of a stimulus when judging self-descriptiveness. Furthermore, given the lesion study described above (Marquine et al., 2016), self-importance information represented in the mPFC might be necessary for accurate and consistent self-knowledge.

Our findings have far-reaching implications. First, they can be contextualized in psychological models of the self-concept. One family of such models depicts the self-concept as an associative network structure where the self is a central entity (node) connected to a number of self-relevant features (e.g., “young,” “university student”) that are themselves connected to each other (Greenwald and Pratkanis, 1984; Kihlstrom and Cantor, 1984).

Researchers further added associative strength to the network model so that some of features (nodes) are more or less strongly connected to the self (and each other; Greenwald et al., 2002). Although these researchers considered the strength of association as "the potential for one concept to activate another" (p. 5), its psychological meaning was unspecified. Our findings suggest that, for links (edges) directly connected to the self, strength of association may be understood as degree of self-importance. Given that associative strength is considered responsible for reaction time facilitation or inhibition during the IAT, our findings generate a hypothesis about reaction time facilitation (e.g., priming effects) based on information self-importance, but not self-descriptiveness (although factors other than self-importance are likely to affect reaction times, such as valence). For example, if being a writer is important to an individual, processing speed for the word "writer" will be facilitated after seeing a prime word "self" (or other highly self-important stimulus). Thus, scores on the self-esteem IAT (Greenwald & Farnham, 2000) might reflect the self-importance information processing function of the mPFC as well as the valence processing function of the reward related network. Largely consistent with this possibility, individual difference in implicit self-esteem as measured by the IAT are independently predicted by activation patterns in the mPFC and those in reward-related brain regions (Izuma et al., 2018).

Second, the findings have implications for psychological research on the link between the self-concept and mental health. For example, it is possible that people who have greater self-complexity (i.e., higher number of, and great differentiation between, self-aspects) are less likely to experience depression, physical illness, and stress in response to aversive events (Linville, 1987), especially when they perceive high control over their self-aspects (McConnell et al., 2005). Similarly, individuals who identify with multiple groups, compared with a single group, report lower stress levels (Binning et al., 2009). Other lines of research point to a link between mental conditions or disabilities and the self-concept. For instance, schizophrenia is associated with changes in self-identity (Conneely et al., 2021), and individuals with autism manifest atypical neural self-representation (Lombardo et al., 2010). How information self-

importance is represented in these patients' brains might shed new light on the nature of mental health, including schizophrenia and autism.

Third, the findings have implications for the long-debated nature of the self among psychologists. One stream of research has emphasised the cognitive properties of the self (Kihlstrom & Klein, 1994; Kihlstrom et al., 2003), characterizing its cognitive structure as complex but ordinary (Greenwald & Banaji, 1989). Another stream of research has emphasised the motivational properties of the self (Kunda, 1990; Sedikides & Strube, 1997), emphasising its uniqueness (Alicke & Sedikides, 2009; Sedikides, 2021). Our findings align with the second empirical stream. If the cognitive representation of the self is unique compared with the cognitive representation of other, this uniqueness lies in motivation (here, attribute self-importance) rather than cognition (here, attribute self-descriptiveness). Moreover, the findings have implications for the long-debated nature of the self among philosophers. Numerous philosophers have cast serious doubts on the mere existence of the self (Baggini, 2011; Hofstadter, 2007; Metzinger, 2009; Midgley, 2014). Here, I countered his viewpoint by providing evidence for the representation of the self in the brain, not only in terms self-descriptiveness, but also in terms of self-importance.

In conclusion, research on the self has a long history in psychology (James, 1890), and the question of “where is the self in the brain?” has attracted keen theoretical and empirical interest in the last two decades (Craik et al., 1999). Although earlier neuroimaging experiments found a robust link between mPFC activation and self-reference processing, what information about the self is processed in the mPFC during a self-reference task has eluded an answer. Our research pinned down the nature of the information about the self that is represented in the mPFC: across two experiments, I demonstrated that information about self-importance (how important a stimulus is to one's self-identity), but not self-descriptiveness, is represented in the mPFC. Put otherwise, the self-concept is represented in the mPFC in terms of self-importance. The mPFC is a neural locus of the self-concept, and this neural system may play a pivotal role in maintaining an accurate and consistent self-concept.

Chapter 3 No Facilitation Effect of Self-Importance on Reaction Time in an Evaluative Priming Paradigm

3.1 Abstract

According to associative network models, the self is a central node connected to other nodes that represent concepts related to the self, and that are themselves connected to each other. Connections between nodes are based on strength, that is, the extent to which they can activate each other. However, the psychological meaning of these connections, or associative links, is unknown. The neuroimaging experiment described in Chapter 2 demonstrated that the self-concept is represented in the mPFC in terms of self-importance. In this preregistered behavioural experiment, I evaluated the hypothesis that strengths of associative links depend on self-importance. I used an evaluative priming task paradigm to test the effect of stimuli's importance to the self-concept in terms of reaction times while controlling for the effects of self-descriptiveness and valence. I did not obtain support for the hypothesis. Instead, contrary to the hypothesis, in an exploratory analysis, I found a priming facilitation effect of self-descriptiveness in the self condition (compared to the other condition). This result is consistent with the possibility that the strength of the associative links in an associative network model of the self represents self-descriptiveness.

3.2 Introduction

Having a sense of self is essential for navigating social interactions (Decety & Sommerville, 2003; Sedikides et al., 2021). The self-concept consists of everything one knows about their personality, such as likes, dislikes, traits, values, beliefs, roles, and ingroups (Sedikides & Gregg, 2003; Sedikides & Spencer, 2007). To understand a multifaceted construct such as the self (Klein & Kihlstrom, 1984; Markus & Wurf, 1987), a researcher must consider not merely the

content, but also the structure of the self-concept (McConnell & Strain, 2011), which is rich and elaborate (Cantor & Kihlstrom, 1984). Researchers have proposed associative network models where the self serves as a central node, linking to other interconnected nodes that represent self-related concepts such as traits, roles, and attitude (Greenwald & Pratkanis, 1984; Kihlstrom & Cantor, 1984). The strength of these links varies depending on how effectively they activate each other (Greenwald et al., 2002). However, the psychological meaning of the links is unclear.

The self-concept is often assessed in terms of self-descriptiveness (Bradley & Mathews, 1983; Kuiper & Derry, 1982; Rogers et al., 1977; Symons & Johnson, 1997), but it can also be assessed in terms of self-importance. Self-descriptiveness is not to be equated with self-importance. For example, "I am human" or "I have 10 fingers" describes most people, but it is unlikely that it is important for their self-identity. Markus (1977) highlighted that although many words describe us, not all are important. Thus, to investigate the self-concept, it is essential to also include a measure of self-importance. In an experiment by Markus (1977; (N = 48), participants were assigned to one of three groups based on their ratings of self-descriptiveness and self-importance along the dependence-independence dimension. The three groups were: (1) those who rated themselves at the extreme end of dependence; (2) those who rated themselves at the extreme end of independence; (3) those with moderate ratings (aschematics). Participants viewed traits for 2 seconds and judged whether the traits were descriptive of themselves. Reaction times were assessed. Independent participants were faster to respond to traits related to independence, dependent participants were faster to respond to traits related to dependence, and aschematics showed no difference in reaction time. A follow-up experiment on gender role orientation produced similar results (Markus et al., 1982). The findings suggest that people need shorter time to judge the relevance of traits that are highly important and descriptive, because these traits are closely linked to the self-node in memory (Kihlstrom & Cantor, 1984). Yet, these experiments (Markus, 1977; Markus et al., 1982) did not differentiate self-importance from self-descriptiveness. The relevance of self-importance was

highlighted by further research. For example, those who consider morality a highly self-important characteristic are more likely to help outgroup members (Reed & Aquino, 2003).

The neuroimaging experiment described in Chapter 2 (i.e., Levorsen et al., 2023) demonstrated that the self-concept is represented in the medial prefrontal cortex (mPFC) in terms of self-importance. While in an fMRI scanner, participants viewed a broad range of stimuli (e.g., physical characteristics, traits, likes, dislikes, roles) and judged if each described them. Participants rated the stimuli on self-importance, self-descriptiveness, and potential confounds (i.e., valence, autobiographical memory.) The findings indicated that the self is represented in the mPFC (and, by implication, the brain) in terms of self-importance rather than self-descriptiveness.

Experiments using implicit measures have reported a facilitation effect when self-related stimuli and positive stimuli are paired compared to control ones. Implicit measures, most prominently the IAT (Greenwald et al., 1998), have also been used in research on self-esteem, an attitude toward the self (Rosenberg, 1979). Greenwald et al. (2002) proposed an associative network model of the self-concept. In this model, self-esteem represents the strength of the link between self and valence. Although the self can be linked to either positive or negative valence, most individuals have positive self-esteem (Greenwald et al., 2002). Greenwald and Farnham (2000) conducted an experiment in which they measured self-esteem with the IAT. Compared to negative words, participants were faster to categorise self-related words with positive words. Similarly, evaluative priming task experiments (Koole & Coenen, 2007; Tao & Zhang, 2012) found faster reaction times in categorising positive words (compared to negative ones) when participants were primed with self-related words or pictures (compared to other-related stimuli). Also, a neuroimaging experiment (Izuma et al., 2018) using fMRI and the IAT reported that activation in reward-related regions of the brain when viewing images of the self predicted individual differences in implicit self-esteem. The collective evidence indicates that the self-node in an associative network model is strongly linked to positivity (Greenwald et al., 2002).

The aim of the current experiment was to find out if links between the self node and self-related concepts represent self-importance. I assumed that associative strength can be operationalised as facilitation in terms of reaction times (Greenwald et al., 2002). I used an evaluative priming task paradigm to test if reaction times for self-related stimuli are faster when preceded by a highly self-important prime word. Participants provided characteristics related to themselves, which I used as stimuli in the priming task. Based on prior findings (Levorsen et al., 2023 as reported in Chapter 2; Markus et al., 1977), I hypothesised a priming facilitation effect of self-importance in the self (versus other) condition.

3.3 Materials and Method

3.3.1 Preregistration

I preregistered hypotheses, sample size, data analyses plan, and exclusion criteria on the Open Science Framework (https://osf.io/v4jzr). Unless otherwise stated, I followed the preregistered analyses plan.

3.3.2 Participants

I recruited 83 University of Southampton undergraduate students, who stated that they were native English speakers and above the age of 18 years. To dissociate the effect of the independent variables on reaction time, I calculated the intercorrelation for each participant's ratings and selected stimuli with the lowest correlation (see "Stimuli Set Preparation" below, for exact steps). If the correlation for a participant's selected stimuli was > .55, I reimbursed them for completing the two online questionnaires but did not invite them to take part in the behavioural experiment session. I excluded 22 participants, because their correlation was too high. Sixty-one participants completed all three parts of the experiment. I then excluded one participant due to a high rate of incorrect responses (> 20%). Also, I excluded another participant, as they were not a native English speaker. The final sample consisted of 59 participants (39 women, 19 men, 1 non-binary) aged 18-38 years (M = 20.50, SD = 3.45). The

experiment was approved by the University of Southampton ethics committee. Participants either received £12.50 or course credit.

3.3.3 Experimental Procedure

The experiment comprised three parts: (1) first online questionnaire; (2) second online questionnaire; (3) behavioural experiment session. The two online questionnaires were similar to those of Levorsen et al. (2023; Chapter 2). The purpose of these two online questionnaires was to create a set of stimuli that covered a wide range of participants' self-concept. On average, the first and second questionnaires were separated by 2.9 days, and the second questionnaire and the behavioural session were separated by 6.4 days.

3.3.3.1 First Online Questionnaire

During the first online questionnaire, participants provided at least 30 self-related items by responding to the prompt "I___" or "My___". They were presented with examples such as physical characteristics (e.g., I am tall), personality (e.g., I am social), likes or dislikes (e.g., food, music, artists), and groups to which they belonged (e.g., university, department, clubs). Participants listed not only strengths and likes, but also weaknesses and dislikes. This task was similar to the Twenty Statements Test (Kuhn & McPartland, 1954). Table 3.1 displays examples of stimuli.

3.3.3.2 Second Online Questionnaire

The second online questionnaire consisted of some items provided by the participants in the first online questionnaire and some items added by the experimenter. The latter were included to dissociate the three dimensions (i.e., self-importance, self-descriptiveness, valence). In total, the second online questionnaire included 80 items. Participants rated each item on self-importance (1 = *not important at all*, 7 = *very important*), self-descriptiveness (1 = *not descriptive at all*, 7 = *very descriptive*), and valence (1 = *very negative*, 7 = *very positive*). I counterbalanced the order of the three dimensions.

3.3.3.3 Stimulus Set Preparation

For each participant, I selected a final stimulus set of 40 items. To dissociate the effect of the dimensions (i.e., self-importance, self-descriptiveness, valence), I aimed to select items with the lowest inter-correlation. In particular, I randomly selected 40 of the 80 items (from the second questionnaire) and calculated the correlation across: (1) self-importance, (2) self-descriptiveness; (3) valence; (4) whether the item was provided by the participant (coded as 1) or the experimenter (coded as 0). Next, I noted the highest correlation ($r_{highest}$) and repeated this process 10,000,000 – 1,000,000,000 times. Finally, I selected 40 items with the lowest $r_{highest}$. Table 3.1 displays examples of stimuli.

Table 3.1 Examples of Items Used in the Experiment

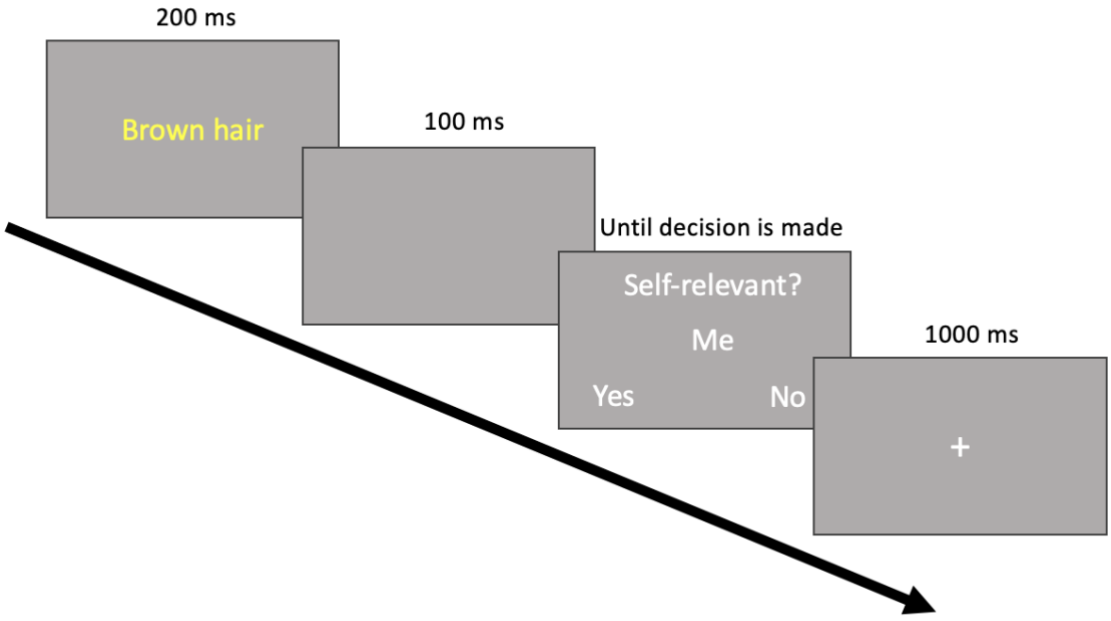
Physical	Social	Attributes	Other
Tall	Psychology Netball	Self-conscious	Paragliding
Fair skin	Islamic society	Optimistic	Divorced parents
Left-handed	Girls grammar school	Scared of heights	Great British Bake Off
Beautiful	Daughter	Feel underappreciated	Taylor Swift
Slightly crooked teeth	Student ambassador	Thoughtful	Gap year
Lean	Irish	High IQ	Part time job
Blue-green eyes	Christian	Short-tempered	Younger siblings

3.3.4 Behavioural Experiment Session

As part of the final behavioural task, I provided participants with task instructions and practice trials. They completed an adapted version of an evaluative priming task (Figure 3.1). They were instructed to ignore the prime item and to quickly and accurately determine whether the target item was self-related (I, My, Me, Mine, Self) or other-related (They, Them, Their, Theirs or Other). Participants completed six blocks, with each block consisting of 40 trials (one item per trial, so that each selected item was presented once per block). Across the six blocks, each item was paired with three self-related targets and three other-related targets. The trial order within each block was randomised. For each trial, participants were first presented with a prime

item in the middle of the screen in yellow font. The prime item remained on the screen for 200 ms and was followed by a blank screen for 100 ms. Next, they were presented with a target item that was either self-related or other-related. The question "Self-relevant?" appeared above the target item. At the bottom of the screen, a "yes" and "no" appeared in counterbalanced order across participants. Participants answered "yes" or "no" with the "f" or "j" button on the keyboard. The target item remained on the screen until participants had responded. Each trial was separated by an inter-trial-interval of 1 second. Overall, each participant completed 240 trials. Demographic questions concluded the experimental session. I used Psychtoolbox (http://psychtoolbox.org/) with MATLAB software (version 2022, http://www.mathworks.co.uk) to program the task.

Figure 3.1. Example of Trials During the Experimental Task



Note. In both conditions participants were presented with a prime word, followed by a blank screen. Next, they judged if a target word was self-relevant. In the self condition the target word was self-relevant (I, My, Me, Mine, Self). In the other condition the target words were other-related (They, Them, Their, Theirs or Other).

3.3.5 Data Analysis

I analysed the data using the Statistics and Machine Learning Toolbox with MATLAB software (version 2019a, http://www.mathworks.co.uk). I ran a repeated measures linear mixed effect model. The fixed effects were condition (self or other), self-importance ratings (1-7 rating), self-descriptiveness ratings (1-7 rating), and valence ratings (1-7 rating). I also included number of characters of target words as a covariate, and participants as a random effect (random intercepts). Further, I tested the interaction effect between condition and each of the other fixed effects. The dependent variable was reaction time. Following standard procedures for reaction time data, I log-transformed the data to correct for skewness (Greenwald et al., 1998).

I hypothesised that participants would respond faster to stimuli rated high (compared to low) on self-importance in the self condition compared to the other condition (i.e., Self-importance × Condition interaction on reaction time). As prior research has found a self-positivity bias in implicit measures (Greenwald & Farnham, 2000; Koole et al., 2001; Koole & Coenen, 2007), valence might affect reaction times. A Valence × Condition interaction is likely, such as participants would respond faster to stimuli rated high (compared to low) on valence in the self condition compared to the other condition.

3.3.5.1 Exploratory Analyses (Not Preregistered)

Given that most participants did not use the full range of the 7-point rating scale for self-importance, self-descriptiveness, and valence, I ran exploratory analyses with standardised rating values. The model remained the same. In addition, I ran a t-test to assess the difference in reaction time between the two conditions (self versus other).

3.3.6 Data Reduction

I discarded trials in which reaction times were less than 0.25 seconds and more than 1.5 seconds (mean reaction time exclusion was 1.92%). I also discarded trials in which participants responded incorrectly (mean error rate was 2.18%).

3.4 Results

The average reaction time was 0.567 seconds (*SD* = 0.177 seconds) in the self condition and 0.596 seconds (*SD* = 0.186 seconds) in the other condition. A paired sample t-tests revealed faster reaction time in the self condition compared to the other condition $t(58) = -6.338$, $p < 0.001$, $M = -0.050$, $SD = 0.061$, 95% CI [-0.066, -0.035], $d = -0.825$.

3.4.1 Linear Mixed-Effects Model

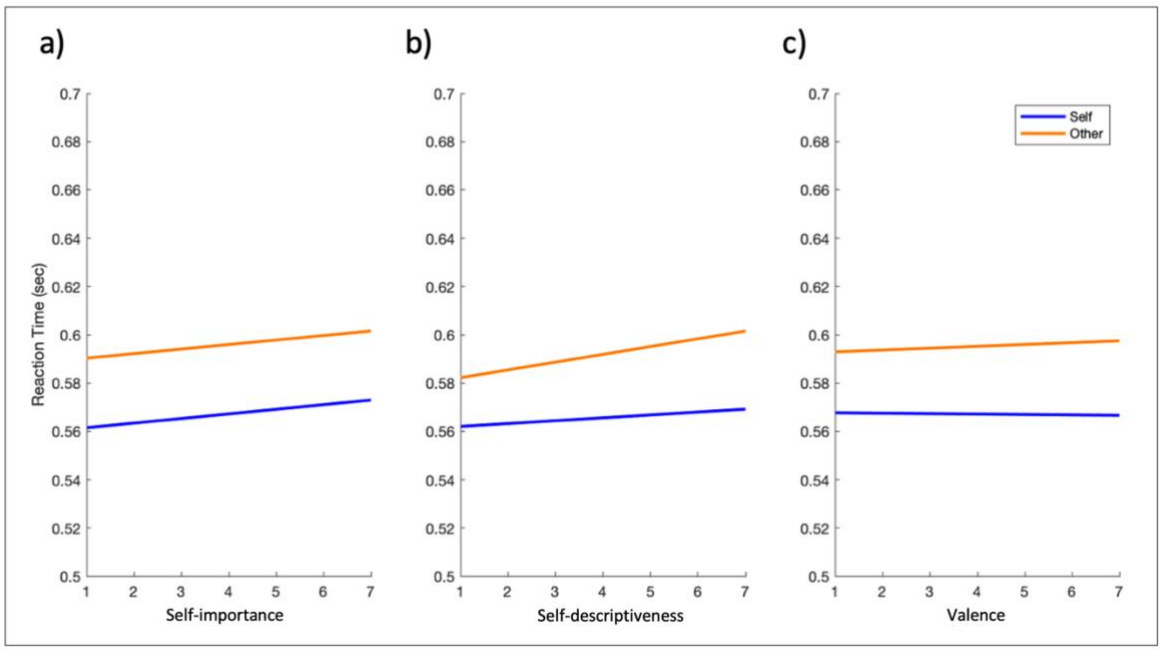
I display results of the model in Table 3.2. Inconsistent with the hypothesis, the Self-Importance × Condition interaction was not significant. Also, inconsistent with self-positivity, the Valence × Condition interaction was not significant. There was a significant effect of number of target characters on reaction times indicating that the longer the target word, the slower the reaction times. No other significant main effect or interaction emerged. I display in Figure 3.2 line graphs demonstrating the association between each rating (self-importance, self-descriptiveness, valence) and reaction times. The random intercept variance for subject was 0.114 (*SD* = 0.246).

Table 3.2 Mixed-Effects Model for Reaction Time Predicted by Condition, Self-Importance, Self-Descriptiveness, Valence, and Number of Target Characters

Fixed effects	Estimate	SE	t	p
Intercept	-0.681	0.022	-31.089	<0.001
Self-importance	-0.001	0.002	-0.499	0.618
Self-descriptiveness	0.003	0.002	1.503	0.133
Valence	-0.001	0.002	-0.241	0.810
Condition	0.027	0.017	1.568	0.117
Target characters	0.024	0.002	11.187	<0.001
Valence × Condition	-0.003	0.003	-1.062	0.288
Self-importance × Condition	0.002	0.003	0.671	0.502
Self-descriptiveness × Condition	-0.003	0.003	-1.258	0.208

Formula: log reaction time ~ 1 + self-importance*condition + self-descriptiveness*condition + valence*condition+ target characters + (1 | subject)

Figure 3.2 Line Graphs Illustrating Reaction Time on the Y-Axis and the Three Ratings on the X-Axis



Note. (a) Self-importance; (b) Self-descriptiveness; (c) Valence. Blue lines represent the self condition and orange lines represent the other condition. Raw reaction times are used for illustration.

3.4.2 Exploratory Analysis with Standardised Ratings (Not Preregistered)

Given that most of the participants did not use the full range of the 7-point rating scale (57.63% for self-importance, 66.10% for self-descriptiveness and 59.32% for valence), I standardised the values and ran an exploratory analysis with standardised rating scores (otherwise the model remained the same as the preregistered model).

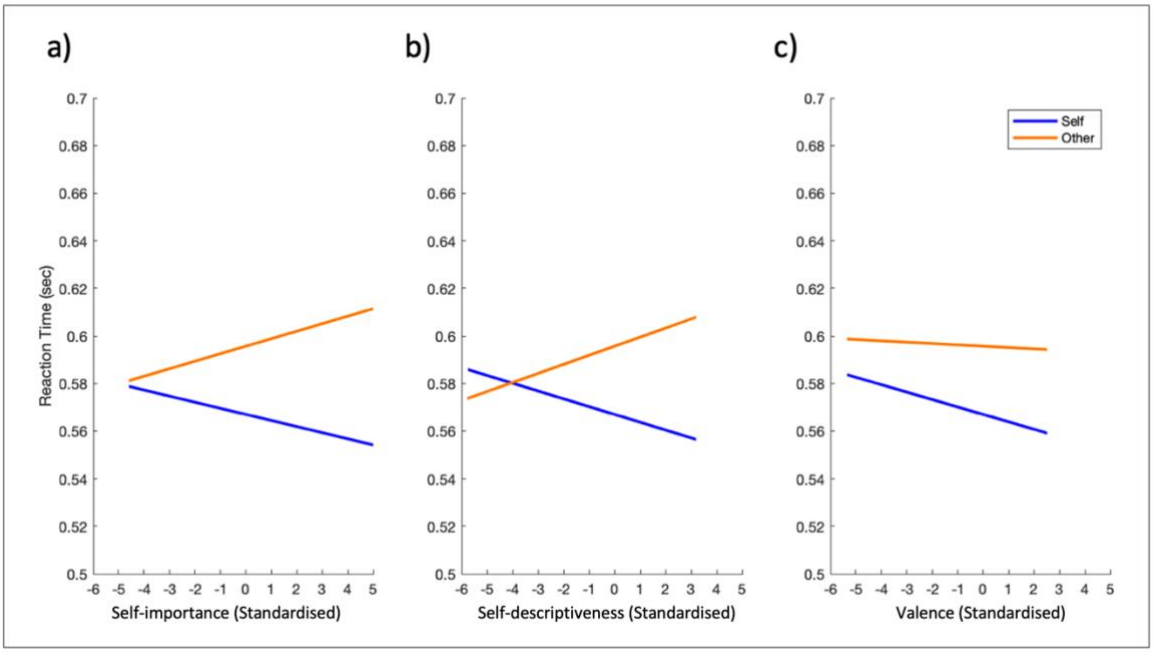
I display the results of the model with standardised rating scales in Table 3.3. The Self-importance × Condition interaction was not significant. However, the Self-descriptiveness × Condition interaction was significant: there was a priming facilitation effect of self-descriptiveness on self-related (compared to other-related) target stimuli. Figure 3.3 displays the association between each rating and reaction time: Figure 3.3a shows the line charts for self-importance, whereas Figure 3.3b shows the line charts for self-descriptiveness. The effect of number of target characters on reaction time was significant. No other significant effects emerged. The random effects remained the same as in the preregistered model.

Table 3.3 Mixed-effect Model for Reaction Time Predicted by Condition, Standardised Self-Importance, Standardised self-Descriptiveness, Standardised Valence, and Number of Target Characters

Fixed effects	Estimate	SE	t	p
Intercept	-0.671	0.018	-36.773	<0.001
Self-importance	0.003	0.003	1.035	0.301
Self-descriptiveness	0.006	0.003	1.715	0.086
Valence	-0.002	0.003	-0.585	0.558
Condition	0.003	0.006	0.439	0.661
Target characters	0.024	0.002	11.214	<0.001
Valence × Condition	-0.003	0.004	-0.589	0.556
Self-importance × Condition	-0.006	0.005	-1.276	0.202
Self-descriptiveness × Condition	-0.009	0.005	-2.018	0.044

Formula: log reaction time ~ 1 + standardised self-importance*condition + standardised self-descriptiveness*condition + standardised valence*condition + target characters + (1 | subject)

Figure 3.3 Line Graphs Illustrating Reaction Time on the Y-Axis and Standardised Ratings on the X-Axis



Note. (a) Self-importance; (b) Self-descriptiveness; (c) Valence. Blue lines represent the self condition and orange lines represent the other condition. Raw reaction times are used for illustrative purposes.

3.5 Discussion

Using an evaluative priming task paradigm, I tested the priming effect of self-importance, self-descriptiveness, and valence in two different conditions (self versus other). Contrary to hypotheses, there was no priming effect of self-importance in the self condition. There was a significant effect of number of target characters on reaction times indicating that the longer the target word, the slower the reaction times. No other significant main effect or interaction effect emerged in our pre-registered model (Figure 3.2). When I ran an exploratory analysis with standardised rating values, I still did not find a priming effect of self-importance (although the interaction was in the expected direction; Figure 3.3a). Instead, I found a priming facilitation effect of self-descriptiveness in the self condition relative to the other condition (Figure 3.3b), suggesting that the strength of associative links in a self memory structure depends on self-descriptiveness.

Two features of the experimental design may explain the absence of the hypothesised effects. First, the strength of the prime words in terms of self-importance and valence might not have been sufficient to trigger the spreading of activation. Researchers have discussed whether only extreme (strong) primes or also the presentation of less extreme (weak) primes result in automatic activation and thus spreading of activation to associatively linked concepts (Bargh et al., 1992; Fazio et al., 1986; Klauer & Musch, 2003). In the pioneering evaluative priming experiment, Fazio et al. argued that only strong attitude prime words result in automatic activation of the prime, and thus facilitation and inhibition priming effects to the target. They maintained that a "weak association" or "nonattitude" is "unlikely to be capable of automatic activation" (p. 236). In contrast, Bargh et al. (1992) argued that weaker (less extreme) prime words could also result in automatic activation and evaluative priming effects. Here, to examine the priming effects of self-importance, self-descriptiveness, and valence, I used words of varying extremity on these three dimensions. According to Fazio et al., the primes might not have been strong enough to result in automatic activation. If the associative links in a network model of the self-concept represent self-importance and valence, it is possible that the prime

words did not result in automatic activation; hence, the primes did not spread activation to related concepts through the associative links.

Second, contrary to the majority of the literature that has used the evaluative priming task to investigate the self, I used self-related stimuli as targets rather than primes (Bach et al., 2009; Koole & Coenen, 2007; Tao & Zhang, 2012). The discrepancy of the valence results between prior work and this experiment might be due to using self-words as targets rather than primes in the current experiment. This point might also explain why I did not find the expected interaction effect of self-importance and condition. In line with Fazio et al. (1986), the self-primes in previous experiments might be strong enough to result in automatic activation and consequently a self-positivity priming effect between the self and targets. In other words, had I used self-words as primes, I could have activated the self-node, and thus measured the associative links between the self-node and related concepts.

Future experiments could implicate self-words as primes to further examine if the associative links in memory represent self-importance. If the primes in the current experiment were not activated due to insufficient strength in self-importance and valence, this could have hindered the ability to measure associative links in self-concept memory. I planned this experiment as a follow up to Chapter 2 (Levorsen et al., 2023). In the self-reference task employed in that research, participants judged stimuli in relation to themselves (and a control condition). I found that the stimuli's self-importance is represented in the mPFC. Similarly to how stimuli are judged in relation to the self in the self-reference task, a self-prime that activates the self-node in an associative memory structure might show a priming facilitation effect of self when target stimuli are high on self-importance. Spreading of activation in an associative network model of self might be unidirectional. That is, self-primes might automatically activate and spread activation to concepts that are highly self-important to the self-concept, but, when these concepts are used as primes, they do not spread activation to self-targets. As in Chapter 2, I intended to gauge the effect of different levels of self-importance. Given that participants make a binary decision in the evaluative priming task, the flipped design

(with self-important words as targets rather than primes) would not have allowed for self-importance to be a continuous variable. However, tweaks in the experimental design with self-primes can still address if the associate links in self-concept memory represent self-importance.

Throughout the years, the evaluative priming task and more broadly, implicit measures, have faced challenges regarding the measure's psychometric properties (Fazio and Olson, 2003; Goodall et al., 2011). Specifically, as noted by Fazio and Olson (2003), the test-retest reliability of the evaluative priming task has varied from "abysmally low" (Bosson et al., 2000) to "moderate levels" (Kawakami & Dovidio, 2001). (p.311). Relatedly, others have criticised the evaluative priming task for the measure's low internal consistency (Banse, 1999, 2001; Bosson et al., 2000). Recently Zayas et al. (2022) compared the reliability of the priming effect in an evaluative priming task with self-related stimuli and showed that although the reliability of the priming effect was low when investigating individual differences, the measure showed high reliability when investigating group-level effects. In other words, given that the current experiment investigated group-level effects, it is likely that the null findings of the current experiment are due to other factor than low reliability, such as the limitations (i.e., the use of weak primes and reversed order of primes and targets) discussed above.

Overall, the observed null results could be due to limitation of the experimental paradigm or assumptions of the theoretical framework. The experimental paradigm used in the current experiment is an adapted evaluative priming task that varies from the original task paradigm (Fazio et al.,1986) in several ways. As mentioned, in the current experiment the order of primes and targets are reversed, also the stimuli are less extreme than in the original study (Fazio et al.,1986). Additionally, the task in the current experiment used continuous, rather than dichotomous independent variables, and each stimulus consisted of a few words rather than just one. Although it is possible that our hypothesis - that associative links in a memory structure of the self represent self-importance - is untrue, given the several deviations, the null findings are likely due to changes in the

paradigm. Future research with different task designs might further test the psychological meaning of the associative links in a memory structure of the self.

The results suggesting that the self is represented in terms of self-descriptiveness are consistent with previous work. Segal et al. (1988) tested if the self is organised in a cognitive memory structure by using an adapted Stroop task paradigm. Participants first viewed 60 positive and negative traits (e.g., trustworthy, selfish). Next, they identified and rated traits that were extremely self-descriptive or highly non-self-descriptive. The traits were later used as primes and targets in the Stroop task where all primes were highly self-descriptive and target words were either neutral or highly self-descriptive. Before they colour-named a target word, participants viewed a prime word. Segal et al. found longer reaction times when primes and targets were both self-descriptive compared to when only the target was self-descriptive. This task was based on the assumption that, when primes and targets are more closely interrelated in a memory structure, activation of the prime automatically activates the meaning of the target, causing interference and thus longer latencies for naming the colour of the target. The authors concluded that self-descriptive concepts are interconnected to a greater extent than non-self-descriptive ones. Although these results do not provide information about direct links from the self node to self-related concept nodes or about the degree of self-descriptiveness (just self-descriptive versus non-self-descriptive items), they suggest that self-related concept nodes are more highly interconnected than non-self-descriptive concept nodes. These findings are consistent with the idea of a self-descriptiveness memory structure. The current results extend their findings, indicating that associations between concepts and the self-node are represented in terms of self-descriptiveness.

In conclusion, prior work suggested that the self is represented as an associative network in memory, comprising links that connect the self-node with self-related concepts. However, the psychological meaning of the links remained unclear. Contrary to my hypothesis, I did not find evidence to suggest that associative links in memory represent self-importance. The preregistered model did not yield further answers either. However, testing an exploratory model

with standardised values indicated that associate links represent self-descriptiveness. Future research that uses self-words as primes might further clarify the meaning of the associative links in the structure of the self.

Chapter 4 Decomposing Cognitive Processes in the mPFC During Self-Thinking[2]

4.1 Abstract

Past cognitive neuroscience research has demonstrated that thinking about both the self and other activate the medial prefrontal cortex (mPFC), a central hub of the default mode network. The mPFC is also implicated in other cognitive processes, such as introspection and autobiographical memory, rendering elusive its exact role during thinking about the self. Specifically, it is unclear whether the same cognitive process explains the common mPFC involvement or distinct processes are responsible for the mPFC activation overlap. In this preregistered functional magnetic resonance imaging (fMRI) experiment with 35 human male and female participants, I investigated whether and to what extent mPFC activation patterns during self-reference judgment could be explained by activation patterns during the tasks of other-reference judgment, introspection, and autobiographical memory. Multi-voxel pattern analysis (MVPA) showed that only in the mPFC, neural responses were both concurrently different and similar across tasks. Furthermore, multiple regression and variance partitioning analyses showed that each task - other-reference, introspection, and memory - uniquely and jointly explained significant variances in mPFC activation during self-reference. These findings suggest that the self-reference task involves multiple cognitive processes shared with other tasks, and the mPFC is the unique place where necessary information is gathered and integrated for judgments based on internally constructed representations.

[2] Levorsen, M., Aoki, R., Sedikides, C., & Izuma, K. (2025). Decomposing cognitive processes in the mPFC during self-thinking. *Journal of Neuroscience*, *45*(22). https://doi.org/10.1523/JNEUROSCI.2378-24.2025

4.2 Introduction

Thinking about the self and expressing who one is to others are fundamental aspects of human experience. The self has fascinated researchers for more than a century (Cooley, 1902; James, 1890). Reflecting this enduring interest, the intricate neural architecture of the self has been a persistent focus of inquiry (Frewen et al., 2020; Wagner et al., 2019). Using neuroimaging methods such as fMRI, experiments have established that the midline structures, the mPFC and posterior cingulate cortex (PCC), are active during the self-reference task in which individuals judge if a presented personality trait or attitudinal statement describes them (Denny et al., 2012; Murray et al., 2012).

Although the robust link between the mPFC and self-reference processing raised the possibility that the primary function of the mPFC is involvement in self-relevant information (Kelley et al., 2002; Northoff, 2016), the mPFC is also involved in thinking about other people (Denny et al., 2012; Murray et al., 2012). Based on these observations, some researchers (Gillihan & Farah, 2005; Legrand & Ruby, 2009) criticized the self-specific view of the mPFC, arguing that some general cognitive processes are common to self- and other-reference processing. For example, inferential processing and memory recall seems to be common to both (Legrand & Ruby, 2009). In other words, the mPFC activation during the self-reference task might not be related to the self specifically, but rather it is a result of general cognitive processes which take place during the self-reference task, as well as during other tasks. Indeed, the mPFC and PCC are also known to be activated by autobiographical memory (Kim, 2012; Martinelli et al., 2013) and by decision-making based on internal or subjective criteria (e.g., moral decision-making; Nakao et al., 2012).

From a broader perspective, the mPFC and PCC are considered the core hubs of the default mode network – a network of brain regions that show heightened activation at rest (Andrews-Hanna et al., 2010). These regions are activated by a variety of tasks that depend on internally constructed representations, including not only self- and other-reference processing,

as well as autobiographical memory, but also introspection (thinking about one's own emotional states), episodic future thinking, creativity, affective decision-making, and spatial navigation (Buckner & DiNicola, 2019; Menon, 2023). For the past two decades, researchers have attempted to identify a key common cognitive process that explains mPFC's involvement in these distinct tasks. However, these attempts are often based on univariate activation overlap (or meta-analyses), and univariate activation overlap does not constitute strong evidence for a common cognitive process across tasks (Levorsen et al., 2023; Woo et al., 2014). Thus, experimental evidence on the extent to which different tasks share a common cognitive process(es) is lacking.

Recently, fMRI experiments using MVPA and representational similarity analysis (RSA) approach have compared patterns of activation for self-reference processing to a few other tasks. For example, Chavez et al. (2017) demonstrated that self-reference processing evoked similar activation patterns in ventral mPFC as positive affect (see also Yankouskaya et al., 2017). When comparing self-reference to other-reference, experiments have demonstrated distinct patterns of activation in mPFC (Courtney & Meyer, 2020; Feng et al., 2018; Koski et al., 2020; Parelman et al., 2022). Although these results begin to clarify scholarly understanding of affective and cognitive processes during self-thinking, the degree to which other internally focused processes (namely, introspection and memory) explain self-reference, remains unknown.

In the present experiment, using RSA and MVPA, I aim to test similarities and differences in neural responses between the self-reference task and three other tasks that also rely on internal representation and are known to robustly activate the mPFC. These are the other-reference, autobiographical memory (Addis et al., 2007; Summerfield et al., 2009) and introspection tasks (Goldberg et al., 2006; Gusnard et al., 2001). Furthermore, using variance partitioning analysis, I aim to quantify how much of explainable variance in mPFC activation patterns during self-thinking can be explained by activation patterns during the other three tasks.

4.3 Materials and Methods

4.3.1 Preregistration

I preregistered the sample size, hypotheses, participant exclusion criteria, and data analysis plan at the Open Science Framework (https://osf.io/mn9fz). Unless otherwise noted, I analysed the data in accord with the preregistration.

4.3.2 Participants

The experiment was approved by the ethics committee of Kochi University of Technology, Japan. Before the online autobiographical memory session, participants ticked a box to indicate their consent. I obtained written consent prior to the fMRI experiment.

The final sample comprised 35 students (8 women, 27 men), ranging in age from 18 to 22 years (M =19.47, SD =1.08). The sample size was based on similar previous experiments (Chavez et al., 2017; Wen et al., 2020; Yankouskaya et al., 2017). I remunerated them with 2,500 Japanese yen. Participants were right-handed, had no history of psychiatric disorders, and had normal or corrected-to-normal vision. I excluded data from one additional participant due to excessive head movement (preregistered exclusion criteria of >3 mm).

4.3.3 Experimental Procedure

The experiment consisted of two parts: (1) online autobiographical memory survey, (2) fMRI experiment. The two sessions took place on separate days, 6.97 days apart on average (SD = 2.54).

4.3.3.1 Online Autobiographical Memory Session

I adapted the autobiographical memory task from (Wen et al., 2020). Prior to the fMRI scan, I instructed participants to write down 15 autobiographical memories, corresponding to one of 15 events each. These memories should pertain to an event bound to a specific time and context that occurred more than one year ago, but after the age of 10 years. The memories ought

to be clear so that participants be in a position to remember the relevant people, objects, and location in detail.

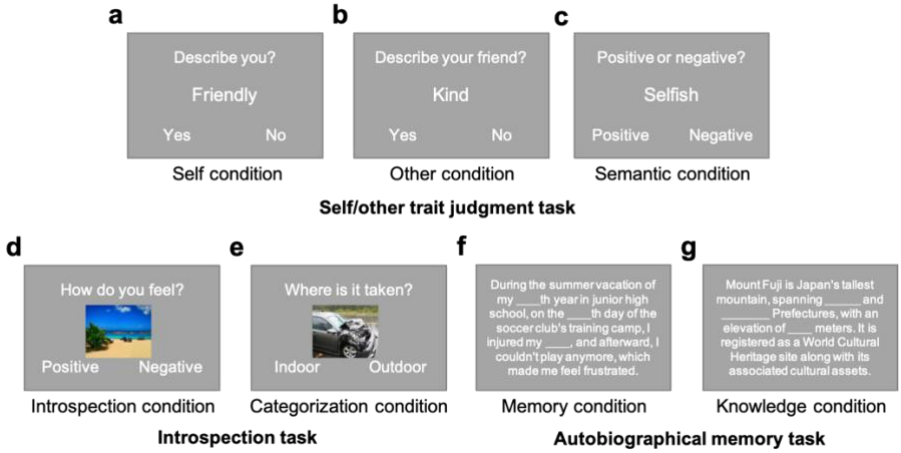
4.3.3.2 Stimuli Preparation

I selected for each participant 10 of the 15 listed event memories. I used the selected memories as stimuli in the autobiographical memory task during the fMRI experiment. I based memory selection on the amount of detail and number of characters included in each description. I matched the number of characters with stimuli in the general knowledge task (see below). For each memory, I removed critical words and replaced them with blank underscores prior to the fMRI experiment.

4.3.3.3 The fMRI Experiment

The fMRI experiment consisted of the following three tasks (Figure 4.1): (1) self/other trait judgement task; (2) introspection task; (3) autobiographical memory task. The self/other trait judgment task had three conditions (Figure 4.1a-c), whereas the introspection (Figure 4.1d & 4.1e) and autobiographical memory (Figure 4.1f & 4.1g) tasks had two conditions each. Thus, there was a total of seven conditions. Participants completed five fMRI runs, with each run lasting approximately 6.5 minutes. Each run included two blocks of seven conditions for a total of 14 blocks. I pseudorandomized the block order within each run, so that the same task block was not presented twice in a row. At the beginning of each block, participants viewed a cue for 1 second indicating that the task that was about to commence. All text stimuli were in Japanese. I programmed all tasks in Psychtoolbox (http://psychtoolbox.org/) with MATLAB software (version 2018a; http://www.mathworks.co.uk).

Figure 4.1 Examples of a Trial/Block for Each of the Seven Conditions Across the Three Tasks



Note. The self/other trait judgment task consisted of (a) self-reference condition, (b) other-reference condition, and (c) semantic condition. The introspection task consisted of (d) introspection task and (e) categorisation task. The autobiographical memory task consisted of (f) memory condition and (g) general knowledge condition.

4.3.3.3.1 Self/Other Trait Judgement Task

The stimuli comprised 40 trait adjectives from a pool of normalized trait adjectives (Anderson, 1968), which I translated into Japanese. The stimuli consisted of an equal number of positive (e.g., “honest,” “trustworthy”) and negative (e.g., “mean,” “greedy”) traits. For each trial, I presented a trait in the middle of the screen. In the self-reference block (Figure 4.1a), we asked participants to judge whether each trait describes them. In the other-reference block (Figure 4.1b), before fMRI scanning, I asked participants to write down the name of one of their close friends on a piece of paper. During scanning, I instructed them to judge whether each trait describes this specific friend. In the semantic judgment block (Figure 4.1c), I instructed them to judge whether each trait is positive or negative. The same 40 adjectives were used across the three tasks. I presented each trial for 2 sec, followed by a 1 second fixation cross, and I presented four traits in each block (12 sec per block). I randomly determined for each participant the order of traits in each of the self-reference, other-reference, and semantic

conditions, but each block always included two positive and two negative words. I presented a fixation cross for 12 sec before the next block.

4.3.3.3.2 Introspection Task

I adapted the introspection task from (Gusnard et al., 2001). It consisted of two conditions: introspection and categorisation. I downloaded 40 picture stimuli (i.e., images of objects, animals, or sceneries) from the Open Affective Standardised Image Set (Kurdi et al., 2017). Half of the stimuli were negative and half positive. For each trial in the introspection block (Figure 4.1d), I presented participants with an image and asked them how the image made them feel. They could respond "positive" or "negative." In the categorisation block (Figure 4.1e), I asked participants to judge whether each picture depicted a scene that was "indoors" or "outdoors." The same 40 images were used across the two tasks. For each participant, the order of images was randomly determined in each of the introspection and categorisation tasks, but each block always included two positive and two negative images. I presented each image for 2 sec and displayed a fixation cross for 1 sec before the next image appeared (each block lasted 12 sec). After an introspection/categorisation block, I displayed a fixation cross for 12 sec before the next block.

4.3.3.3.3 Autobiographical Memory Task

The autobiographical memory task (Wen et al., 2020) comprised two conditions: memory and knowledge. For each trial in the memory condition (Figure 4.1f), participants encountered one of the memories they had previously listed in the online autobiographical memory session. Each memory consisted of, on average, 67.6 Japanese characters (*SD* = 7.25), which I matched with the length of the stimuli used in the knowledge condition. Within each memory, I replaced three critical words with blank underscores. I asked participants to recall the memory and fill in the blanks for the missing words, but do so in their mind rather than by pressing a button (i.e., I recorded no responses during this task).

In the knowledge condition (Figure 4.1g), I presented participants with text related to general knowledge (*M* = 67.8 characters, *SD* = 8.11 characters), such as a description of a common topic (e.g., Mt. Fuji, football, seatbelt), in which I replaced certain words with blank underscores. I instructed participants to think of appropriate words to fill in the blanks.

In both conditions, I presented each text stimulus for 14 sec and followed it by a fixation cross (4-6 sec). Next, I asked: "Were you recollecting a specific event?" (1 = *not at all*, 5 = *extremely vividly*). Participants had up to 6 sec to respond. I presented a fixation cross for 10 sec before the next block.

4.3.4 Behavioural Data Analysis

I conducted a one-way Analysis of Variance to compare reaction time (RT) and response rates across the self-reference, other-reference, and semantic judgment tasks. Given that the RT data were not normally distributed, I log-transformed them beforehand. I followed up significant effects with a Bonferroni-corrected tests. All reported p values were two-sided.

4.3.5 fMRI Data Acquisition

I acquired images using a 3.0 T Prisma Siemens MRI scanner with a 64-channel phased-array head coil. For functional imaging, I used T2*-weighted gradient-echo echo-planar imaging (EPI) sequences. I acquired 42 contiguous transaxial slices (covering almost the entire cerebrum) with a thickness of 3 mm, in an interleaved order. I acquired the images with the following parameters: Time repetition (TR) = 2500 ms, echo time (TE) = 25 ms, flip angle (FA) = 90°, field of view (FOV) = 192 mm^2, matrix = 64 × 64. Additionally, I acquired a T1-weighted structural image (with 1mm isotropic resolution) from each participant.

4.3.6 fMRI Data Preprocessing

I carried out preprocessing and statistical analysis in SPM 12 (Welcome Department of Imaging Neuroscience), implemented in MATLAB (Math Works). To allow for T1 equilibration, I discarded the first four volumes before preprocessing and data analyses. I used SPM 12's

preproc_fmri.m script to perform preprocessing of the fMRI data. I spatially realigned all functional images within each run to the mean using 7th-degree B-spline interpolation. I normalized the volumes to MNI space using a transformation matrix that I obtained from the EPI normalization of the first participant to the EPI template. I resampled the volumes to a voxel size of 3 × 3 × 3 mm^3, that is, I retained the original voxel size. I used the 7th-degree B-spline interpolation option for normalization. I applied spatial smoothing (of 8 mm FWHM) to the data for the whole brain univariate analysis. To maintain fine-grained activation patterns, I did not apply smoothing to the data for representational similarity analysis nor for multivariate pattern analysis.

4.3.7 Univariate fMRI Analysis

4.3.7.1 General Linear Model (GLM)

I first ran a conventional GLM analysis, modelling separately each of the seven task blocks (i.e., conditions) with duration of 12 sec, except for the autobiographical memory and general knowledge tasks which had a duration of 14 sec. The memory and knowledge tasks had a rating phase which I modelled separately as a nuisance regressor (duration = participant's response time). I also included six head motion parameters as nuisance regressors. To examine mPFC activation, I created the following six contrast images for each participant: (1) self > semantic; (2) other > semantic; (3) self > other; (4) introspection > categorisation; (5) memory > knowledge; (6) rest > semantic + knowledge + categorisation. I used the last contrast to identify regions that showed increased activations during passive rest compared to externally focused tasks (Gusnard et al., 2001; Shulman et al., 1997; Wen et al., 2020). Furthermore, I created the seven additional contrast images (each of the seven tasks relative to the implicit baseline [i.e., rest]). The spmT images from these contrasts were used in the subsequent MVPA and RSA analyses (details below).

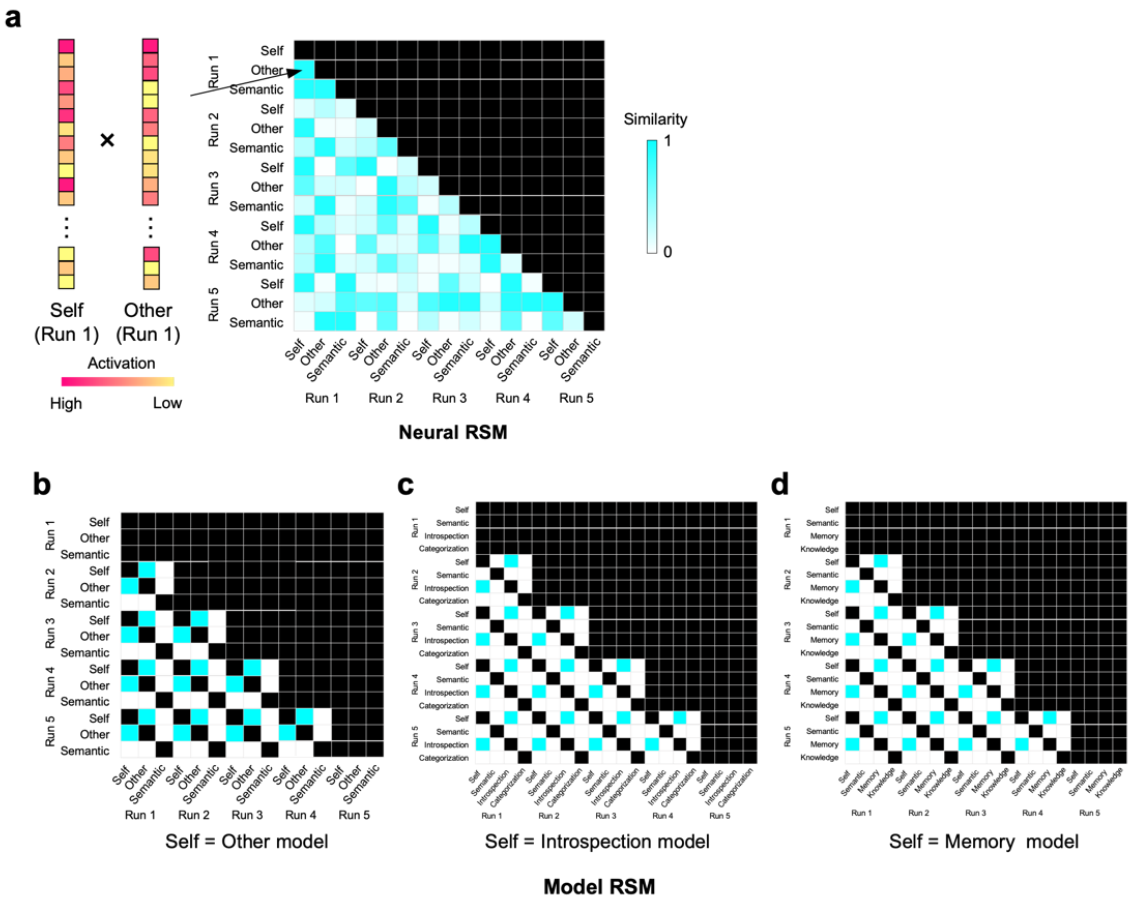
4.3.7.2 Group Analysis

I conducted a second-level whole brain group analysis for each of the contrasts. I set the statistical threshold at $p < 0.001$ voxel-wise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple comparisons).

4.3.8 Representational Similarity Analysis

I conducted the RSA to test the similarity in activation patterns between the self and each of the other-reference, introspection, and memory conditions. For each participant, neural data were extracted from the spmT image of each contrast, and I computed neural representational similarity matrix (RSM; Figure 4.2a) based on Pearson correlation across activation patterns in each pair of conditions across the five runs. There are three model RSMs (Figure 4.2b-d), each of which addresses the similarity between the self and (1) other, (2) introspection, (3) memory. Given that I are interested in the similarity between the self and other, independently of similarities across the remaining conditions, I excluded from analyses the irrelevant conditions. For example, when testing the self = introspection model (Figure 4.2c), I excluded the other-reference, memory, and knowledge conditions so that pattern similarities involving those irrelevant conditions would not affect the results. I evaluated the fit between the neural RSM and model RSM via Kendall's tau-a for each participant (Nili et al., 2014). Activations of any two conditions within the same run are likely to be positively correlated largely due to shared physiological noises (Alink et al., 2015); as such, I excluded correlations between any pairs of conditions within the same run to the model RSM. I also excluded correlations between neural responses of the same conditions (Ritchie et al., 2017). I ran these RSAs using a searchlight approach (explained below).

Figure 4.2 Schematic Illustrations of Representational Similarity Analysis (RSA)



Note. (a) For each participant, I created a neural representational similarity matrix (RSM) by computing Pearson correlations between activation patterns during two tasks across five runs. (b) Self = Other model RSM. (c) Self = Introspection model RSM. (d) self = memory model RSM. In each neural/model RSM, I excluded cells in black from the analysis. In panels b-d, cells in cyan represent 1 (similar), whereas cells in white represent 0 (dissimilar). I evaluated fit between the neural versus each model RSMs through Kendall's tau-a (Nili et al., 2014).

4.3.9 Classifier-based MVPA

The above RSA tests whether activation patterns are similar between two conditions. I proceeded to conduct classifier-based MVPA to examine whether activations patterns in the two conditions were distinct. I implemented a linear support vector machine (SVM), carried out via MATLAB in combination with LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) (Levorsen et al., 2021; Wake & Izuma, 2017), with a cost parameter of c = 1 (default).

I used MVPA to find out if the activation patterns for the following contrasts were distinct: (1) self > semantic versus other > semantic; (2) self > semantic versus introspection > categorisation; (3) self > semantic versus memory > knowledge. For each participant, neural data were extracted from the spmT image of each of these contrasts. To evaluate classification performance, I employed a leave-one-run-out cross-validation procedure. Thus, I first left out one run in each cross-validation, and, using the data from the rest of runs, I trained a classifier that discriminates (e.g., activation patterns between self > semantic versus introspection > categorisation contrasts). Subsequently, I tested the classifier performance using the data from the left-out run. I repeated this procedure five times leaving out a different run each time, and I averaged the five classification accuracy values. Like the RSA, I ran the classifier-based MVPA using a searchlight approach (below).

4.3.10 Searchlight Analysis

I conducted the RSA and MVPA with a searchlight approach (Kriegeskorte et al., 2006). For the RSA, I extracted local patterns of neural activity from searchlights with a three-voxel radius, so that each searchlight consisted of a maximum of 123 voxels (and less on the edges of the brain). I made a neural RSM from each searchlight and computed Kendall's tau-a between neural versus each of the three model RSMs (Figure 4.2), which I saved for a center voxel, resulting in three correlation maps for each participant.

Similarly, for the classifier-based MVPA, I carried out MVPA within each searchlight, and I saved a classification accuracy for a center voxel, resulting in a total of three classification accuracy maps for each participant. Within each searchlight, I removed mean activity by subtracting the mean value of a searchlight sphere from values of the individual voxel so that mean activation difference across conditions could not account for MVPA results.

4.3.10.1 Group Analysis

I applied smoothing before the group analysis of the RSA and MVPA outputs (with a Gaussian kernel of 4-mm FWHM). Following the smoothing, I entered the Kendall's tau-a maps

and classification accuracy maps into a second-level permutation-based analysis (with 5,000 permutations). I used the Statistical Non-Parametric Mapping toolbox for SPM (Nichols & Holmes, 2002). Within the preregistered mPFC region of interest (ROI), I set a statistical threshold (i.e., voxel-level) at $p < 0.005$, and a cluster-level threshold at $p < 0.05$ (FWE corrected). Outside of the mPFC, I set a statistical threshold at $p < 0.001$, and a cluster-level threshold at $p < 0.05$ (FWE corrected).

4.3.11 ROI Analysis

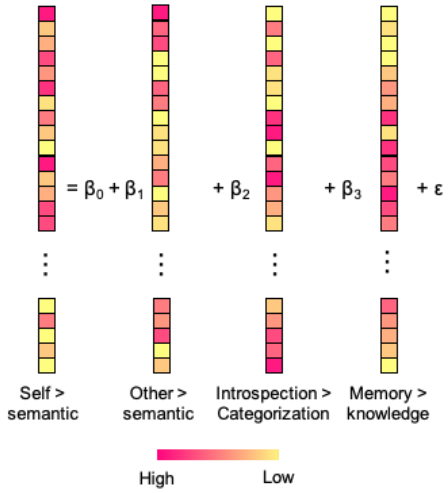
I further investigated the role of the mPFC in thinking about the self by running a ROI analysis. I used Neurosynth (https://neurosynth.org/; Yarkoni et al., 2011) to define our mPFC ROI independently of our data. I downloaded an association map (thresholded at $q < .01$, False Discovery Rate corrected), which I generated from a term-based meta-analysis with the label "self-referential" (downloaded on October 10th, 2023). The mPFC ROI included 308 voxels. I ran the following multivariate pattern regression analyses within the ROI.

4.3.12 Multivariate Pattern Regression

The above RSA and classifier-based MVPA address neural pattern similarity and difference separately for each pair of tasks. I conducted a multivariate pattern regression analysis to compare pattern similarity across multiple tasks within the same framework. I ran a multiple regression analysis where activation patterns of the self > semantic contrast were a dependent variable, whereas those of (2) the other > semantic, (2) introspection > categorisation, and (3) memory > knowledge contrasts were independent variables (Figure 4.3).

As stated above, given that activation patterns of any two conditions within the same run are likely to be positively correlated likely due to shared physiological noise, I ran the regression analysis 20 times (all possible pairs of five runs excluding pairs from the same run) so that independent and dependent variables were always from two different runs. I averaged all outputs (beta values and adjusted R^2) across the 20 regression analyses within each participant.

Figure 4.3 Multivariate Pattern Regression



Note. Activation patterns of the self > semantic contrast were a dependent variable, whereas activation patterns of the other three contrast were independent variables. Independent and dependent variables were always from different runs.

4.3.12.1 Noise Ceiling Model

To provide an estimate of how much systematic variation in activation patterns of the self > semantic contrast could be explained in the data given measurement noise, I included a noise-ceiling model. This model simply included the data from the self > semantic contrast as both a dependent and independent variable (although they were from different runs) in the multivariate pattern regression. Thus, the only difference between the noise ceiling model and original full model (illustrated in Figure 4.3) was the inclusion of activation patterns of the self > semantic contrast as another independent variable in the noise ceiling model.

4.3.12.2 Variance Partitioning Analysis

Following the multivariate pattern regression analysis, I carried out variance partitioning analysis to infer the amount of unique and shared variance between three different predictors. I conducted seven multiple regression analyses: one with all three independent variables as predictors (illustrated in Figure 4.3), three with different pairs of two independent variables as predictors, and three with individual independent variables as predictors. Comparing the

explained variance (R^2) of a model used alone with the explained variance when used with other models would allow us to infer the amount of unique and shared variance between different predictors.

4.3.12.3 Permutation Test

To assess the significance of the findings from the multivariate pattern regression analyses and variance partitioning analysis, I ran permutation tests where voxels were randomly shuffled. The self > semantic contrast and other > semantic contrast have the semantic condition as a common control condition, and this common control condition is likely to bias a beta value associated with the other > semantic activation patterns to a positive direction. Thus, our permutation test randomly shuffled beta activation map of the self-reference condition (i.e., self > implicit rest contrast). I computed a randomly-shuffled-self > semantic contrast image (and a corresponding t-statistics map) so that the effect of the similarity in neural responses between the semantic task versus each of the remaining five tasks remained intact in each permutation. I repeated this step 1,000 times to estimate null distributions. Furthermore, shuffling voxels may overly destroy spatial autocorrelation in the original data, which might bias results of the permutation test. Thus, I smoothed shuffled data via a Gaussian kernel with the standard deviation of 0.86 before conducting a multiple regression analysis (see Burt et al., 2020, for a similar approach). I selected a standard deviation of 0.86, because it produced the smallest sum of square error between the smoothness (quantified as Moran's I based on an inverse Euclidean distance matrix; Moran, 1950) of the original data versus that of shuffled-and-then-smoothed data (repeated 1,000 times; I tried all standard deviation values ranging from 0 to 2.0 with an increment of 0.2).

4.3.13 Deviations from Preregistration

I deviated from the preregistration as follows. First, I preregistered and conducted MVPA testing pattern generalizability (i.e., cross-task classification) which, like the RSA, aims to examine the similarity in activation patterns between two conditions. However, I do not report

relevant results, because they were similar to the results of the RSA reported below; also, this analysis is inappropriate when testing the similarity between the self- versus other-reference conditions due to their common control condition. Second, I did not preregister the following: behavioural data analyses, reaction time (RT)-controlled MVPA, multivariate pattern regression, and variance partitioning analyses.

4.4 Results

4.4.1 Behavioural Results

During the self/other trait judgment conditions, participants pressed one of the two keys in almost all trials in the self (99.6%), other (99.9%), and semantic (99.9%) conditions. There was a significant difference in RT across the three conditions ($F_{(2,68)} = 18.31$, $p < 0.001$). Pairwise t-tests revealed that RTs were significantly different from each other across conditions. RTs in the self-reference condition ($M = 1.21$ sec, $SD = 0.18$ sec) were significantly longer than those in the other-reference condition ($M = 1.14$ sec, $SD = 0.25$ sec; $p_{corrected} = 0.004$) and in the semantic condition ($M = 1.08$ sec, $SD = 0.19$ sec; $p_{corrected} < 0.001$). RTs in the other condition were significantly longer than those in the semantic condition ($p_{corrected} = 0.026$).

I next examined if RTs in the other-reference condition were influenced by response similarity between the self and other, as reported in a previous experiment (Thornton and Mitchell, 2018). I ran a multiple regression analysis with RT in the other-reference condition as a dependent variable, and response similarity as an independent variable (1 = same responses to the same trait, -1 = different responses). I also entered as independent variables participant response (1 = yes, -1 = no), trait valence (1 = positive, -1 = negative), number of characters of each word stimulus, and the interaction between participant response and trait valence. I obtained a significant effect of response similarity ($t_{(34)} = 3.79$, $p = 0.003$). RTs were shorter when the self- and other-reference judgments for the same trait were identical (i.e., both yes or both no). Although this result suggests egocentric anchoring and adjustment in other-reference judgment, I observed a similar effect in the self-reference condition (see below). Number of

characters was significantly related to RTs, meaning the more characters a word had, the slower the participant responded ($t_{(34)}$ = 3.80, p = 0.003). I also obtained a significant Participant Response × Trait Valence interaction ($t_{(34)}$ = 4.42, p = 0.005). Participants were slower to respond yes than no when judging if a negative trait described their friend, whereas they did not differ in their responses to positive traits. No other significant effect emerged.

I conducted the same regression analysis for the self-reference condition to test if RTs in the self-reference condition were influenced by response similarity between the self and other. I found only a significant effect of number of characters ($t_{(34)}$ = 2.97, p = 0.027). The effect for response similarity was trending ($t_{(34)}$ = 2.51, p = 0.086). When I compared beta values for the self- versus other-reference conditions, I observed no significant difference ($t_{(34)}$ = 1.03, uncorrected p = 0.31), suggesting that the significant effect of the response similarity obtained in the other-reference condition might be at least partially explained by unknown stimulus features.

Consistent with prior research (Moran et al., 2006), participants were more likely to endorse a positive trait as self-descriptive and a negative trait as not self-descriptive ($t_{(34)}$ = 4.28, p < 0.001). However, I observed this positivity bias in the other-reference condition as well ($t_{(34)}$ = 8.46, p < 0.001); indeed, this bias was stronger for the other-reference than the self-reference condition, indicating that participants were more other-enhancing than self-enhancing ($t_{(34)}$ = 3.48, p = 0.001). These results are largely consistent with some findings suggesting that self-enhancement is weaker for East Asian compared to Western individuals (Heine & Hamamura, 2007; but see Cai et al., 2016).

During the introspection task, participants pressed one of the two keys in almost all trials in the introspection (99.9%) and categorisation (99.7%) conditions. RTs during the introspection condition (M = 1.08 sec, SD = 0.21) were significantly slower than those during the categorisation (M = 1.14 sec, SD = 0.17) condition ($t_{(34)}$ = 3.79, p < 0.001), likely because some pictures were ambiguous as to whether they were taken indoors or outdoors.

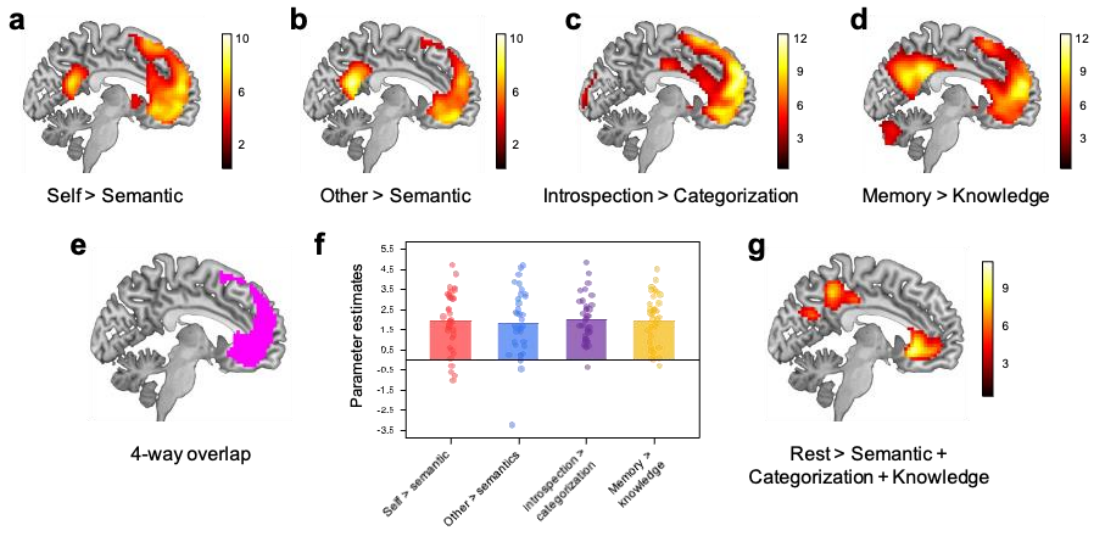
During the autobiographical memory task, participants successfully gave their vividness rating within the time limit of 6 sec for almost all trials in the memory (99.0%) and knowledge (98.7%) conditions. Vividness ratings were significantly higher in the memory (M = 4.37, SD = 0.40) compared to the knowledge (M = 2.45, SD = 0.79) condition ($t_{(34)}$ = 17.73, p < 0.001), testifying to the effectiveness of our memory manipulation.

4.4.2 fMRI Results

4.4.2.1 Univariate Analysis Results

Replicating findings from several experiments (Denny et al., 2012; Murray et al., 2012), the self > semantic contrast significantly activated the midline structure including mPFC and PCC (Figure 4.4a). The other > semantic contrast activated similar regions (Figure 4.4b). Left temporoparietal junction (TPJ) was also commonly activated by the self and other conditions. Although there were some regions that were uniquely activated by either the self > semantic or other > semantic contrast when the self and other conditions were directly compared, the self > other contrast did not lead to any significant activation. The opposite contrast (other > self) revealed only one significant cluster in PCC (303 voxels; x = 6, y = -64, z = 29).

Figure 4.4 Results of the Univariate Analyses



Note. Sagittal slices (x = -6) showing results of the Univariate Analyses (a) Areas significantly activated by the self > semantic contrast. (b) Areas significantly activated by the other > semantic contrast. (c) Areas significantly activated by the introspection > categorisation contrast. (d) Areas significantly activated by the memory > knowledge contrast. (e) Areas commonly activated by the all four contrasts (1,565 voxels). Only mPFC showed significant 4-way overlap. For display purposes, I set statistical threshold at $p < 0.005$ and cluster-$p < 0.05$ (FWE corrected). (f) Parameter estimates of the four contrasts within the mPFC areas commonly activated by the four contrasts (panel e). (g) Areas significantly activated by the rest > semantic + categorisation + knowledge contrast.

As per prior experiments (Goldberg et al., 2006; Gusnard et al., 2001), the introspection > categorisation contrast significantly activated the mPFC (Figure 4.4c). Other activated areas included anterior cingulate cortex, temporal pole, lateral temporal cortex, and lateral occipital cortex.

The memory versus knowledge contrast significantly activated regions previously implicated in autographical memory including the mPFC, PCC/precuneus, posterior inferior parietal lobule (pIPL), and lateral temporal cortex (LTC; Kim, 2012; Martinelli et al., 2013) (Figure 4.4d).

Taken together, the above four contrasts all significantly activated the common region within the mPFC (1,565 voxels; Figure 4.4e). Bilateral temporal poles were also commonly activated by all four contrasts (left x = -36, y = 17, z = -22, 109 voxels; right x = 30, y = 14, z = -22,

185 voxels). No other region was commonly activated. Yet, although the introspection > categorisation contrast activated the PCC (92 voxels), it did not pass our preregistered cluster-level threshold. When I directly compared the four contrasts to each other within the commonly activated mPFC areas, no significant difference emerged ($F_{(2.38, 80.86)} = 0.08$, $p = 0.94$; Figure 4.4f).

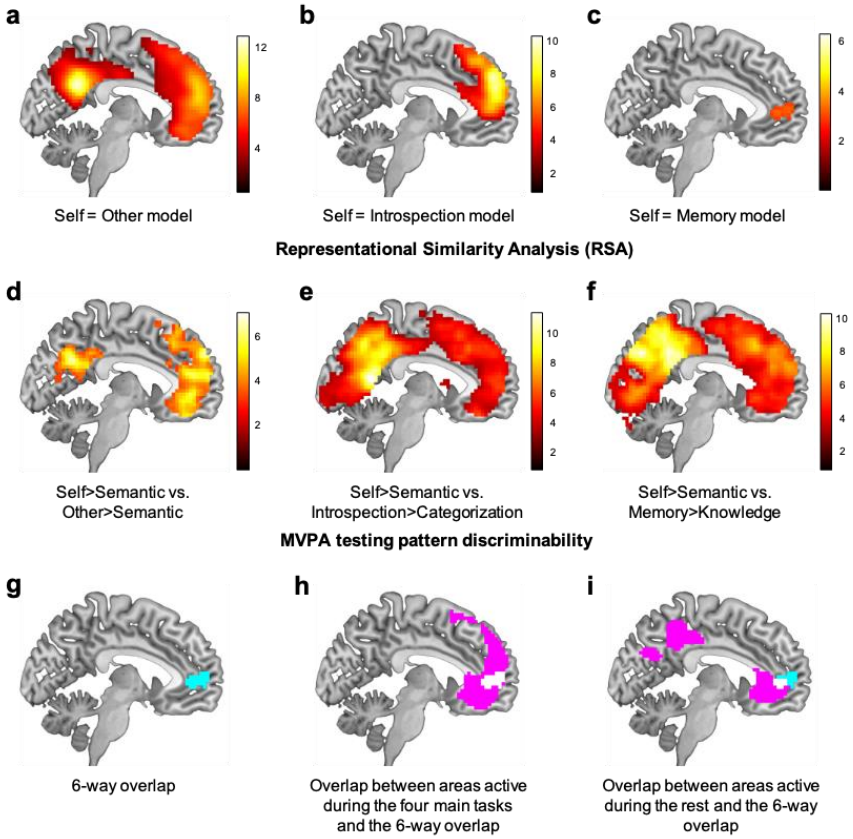
Consistent with previous findings (Gusnard et al., 2001; Shulman et al., 1997; Wen et al., 2020), the rest > semantic + categorisation + knowledge contrast revealed that areas in the default mode network, including mPFC, PCC, IPL, TPJ/AG, and LTC, were active during rest compared to the externally focused tasks (Figure 4.4g).

4.4.2.2 Results of RSA: Are Activation Patterns Evoked by Two Tasks Similar?

The RSA (Figure 4.2) aims to test whether the self-reference judgment evoked similar activation patterns with each of the other-reference judgment, introspection, and autobiographical memory.

The Self = Other model (Figure 4.2b) was significantly associated with a network of brain regions involved in self-reference and social cognition including mPFC, PCC/precuneus, bilateral inferior frontal gyrus (IFG), bilateral superior temporal sulcus, and bilateral temporal pole (Figure 4.5a). However, the other two models were associated only with mPFC and left IFG. In particular, the Self = Introspection model (Figure 4.2c) was significantly associated with mPFC (x = 0, y = 53, z = 35, 1,930 voxels; see Figure 4.5b) and left IFG (extending to temporal pole; x = -51, y = 20, z = 2, 379 voxels). Further, the Self = Memory model was significantly associated with mPFC (x = -12, y = 44, z = 5, 103 voxels; Figure 4.5c) and left IFG (x = -45, y = 26, z = -7, 368 voxels).

Figure 4.5 Results from the RSA, MVPA and Overlaps Across Analyses



Note. (a-c) Sagittal slices (x = -6) showing results from the RSA. Significant areas indicate that activation patterns of the two contrasts were similar. (d-f) Sagittal slices (x = -6) showing results from the MVPA testing pattern discriminability. Significant areas indicate that activation patterns of the two contrasts were distinguishable. For display purposes, I set statistical threshold at $p < 0.005$ and cluster-$p < 0.05$ (FWE corrected). (g) A sagittal slice (x = -6) showing the mPFC area that showed 6-way overlap (overlap across areas shown in panel a-f). (h) A sagittal slice (x = -6) showing overlap between univariate and MVPA results. Magenta represents areas activated commonly by the four univariate contrasts (Figure 4.4e), and white represents 6-way overlapped region depicted in panel g. (i) A sagittal slice (x = -6) showing overlap (white areas) between areas activated by the rest > semantic + categorisation + knowledge contrast (magenta; Figure 4.4g) and the 6-way overlapped region depicted in panel g (cyan).

4.4.2.3 Results of MVPA Testing Pattern Discriminability: Are Activation Patterns Evoked by Two Tasks Distinguishable?

The classifier-based MVPA tested pattern discriminability with a searchlight approach. It addressed whether activation patterns evoked by different tasks were distinguishable or linked to different cognitive processes. Indeed, activation patterns evoked during the self-reference task (relative to the semantic task) were distinguishable from the other-reference task in the mPFC, PCC, and right superior temporal sulcus (extending to the temporal pole; Figure 4.5d). These areas largely overlapped with the areas activated by the self- and other-reference condition relative to the semantic condition (Figure 4.4a and 4.4b), indicating that those areas were commonly activated both by the self and other conditions compared to the semantic condition, but their activation patterns were systematically different. Given that the self and other conditions had the semantic condition as common control, the difference between the self and other conditions is likely to be underestimated in this analysis.

In contrast, activation patterns elicited by the self-reference condition were distinguishable from each of the introspection and memory conditions in a number of regions across the whole brain including the mPFC, PCC, intraparietal lobule, middle temporal gyrus, and TPJ (Figures 4.5e and 4.5f).

These results, together with the RSA results reported above, indicate that mPFC activation patterns during the self-reference judgement were similar to those elicited during each of the other-reference judgement, introspection, and autobiographical memory (Figure 4.5a-c). Nonetheless, they were still distinguishable from activation patterns of each of the three tasks (Figure 4.5d-f). In fact, there was one cluster within the mPFC (Figure 4.5g; a total of 96 voxels) showing significant association/classification accuracy in all six analyses (Figure 4.5a-f), and the mPFC cluster is the only region that showed the 6-way overlap with the cluster size larger than 20 voxels. This 6-way overlap was located in the anterior part of the mPFC (Brodmann Area [BA] 10) and pregenual anterior cingulate cortex (pgACC; BA 32). Furthermore, this cluster

entirely overlapped with the areas commonly activated by the four contrasts in the univariate analyses (Figure 4.5e). It also showed a substantial overlap (53 out of 96 voxels [55.2%]) with the areas significantly active during the rest (Figure 4.5g), indicating that most of the 6-way overlap area (Figure 4.5e) is located in the mPFC within the default mode network.

4.4.2.4 Results of ROI Analysis

I conducted additional ROI analyses to refine the findings and run control analyses. I defined the mPFC ROI independently of our own data using Neurosynth (see Methods). This ROI analysis focused on areas within the mPFC most strongly associated with self-reference processing (Figure 4.6a).

4.4.2.4.1 Does the Difference in Activation Patterns Between the Self- and Other-Reference Conditions Simply Reflect the Difference in RTs Between Them?

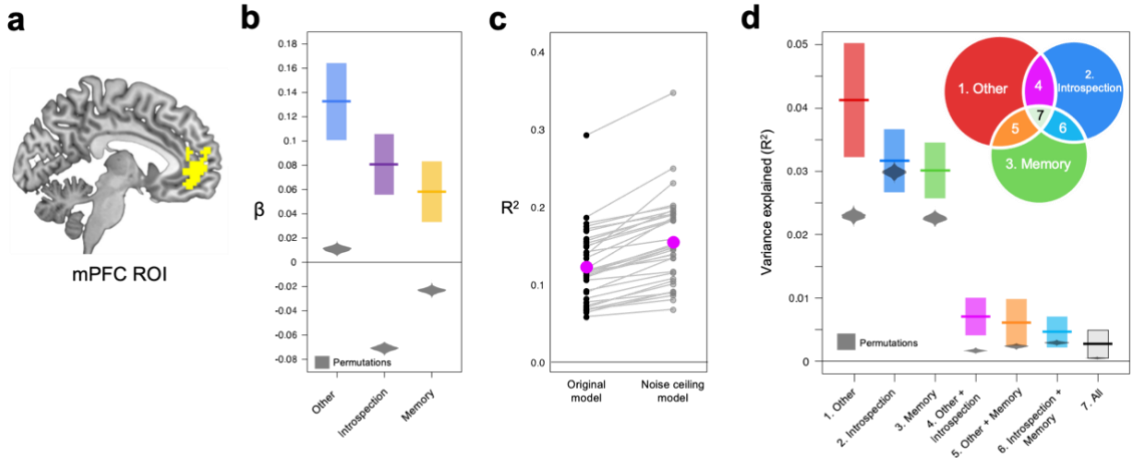
According to our behavioural results, RTs were significantly longer for the self-reference condition compared to the other-reference condition. Thus, the difference in activation patterns between the two conditions might be explained by the difference in RTs (e.g., task difficulty). To rule out this possibility, I ran additional GLM where I categorised self- and other-reference task blocks into short and long RT blocks based on average RTs in each block. I modelled the other five tasks in the same way as the original GLM. Then, I ran an MVPA analysis testing whether it can distinguish activation patterns of the mPFC ROI during the short versus long RT blocks.

Within the mPFC ROI, the average accuracy for classifying the short and long RT blocks was 51.71%, which did not differ significantly from the theoretical chance level of 50% (Wilcoxon signed rank test, $p = 0.31$). Also, it was significantly lower than the accuracy for classifying actual self- versus other-reference blocks (average = 63.14%; paired-sample Wilcoxon signed rank test, $p = 0.002$). I additionally ran the same MVPA (short versus long RT blocks) across the whole brain with a searchlight approach, but did not find any significant area. Taken together, the difference in RTs between the self and other conditions is unlikely to explain the difference in activation patterns between the two conditions.

4.4.2.4.2 Which Task Best Explains Activation Patterns of the Self-Reference Condition?

The results of the RSA reported above (Figure 4.5a-c) indicate that mPFC activation patterns during the self-reference condition were similar to those of the other-reference, introspection, and memory conditions. However, these analyses addressed neural pattern similarity separately for each pair of tasks. To compare pattern similarity across three tasks within the same framework, I carried out a multivariate pattern regression analysis where activation patterns of the self > semantic contrast were a dependent variable, whereas those of the other > semantic, introspection > categorisation, and memory > knowledge contrasts were independent variables (Figure 4.3). Activation patterns of each of the three contrasts were significantly associated with mPFC ROI activation patterns of the self > semantic contrast (Figure 4.6b; all $p_{perm} < 0.001$), suggesting that the similarity in mPFC neural responses between the self-reference task and each of the other three tasks remain significant even after controlling for the effect of neural responses during the other two tasks.

Figure 4.6 Results of the Multivariate Pattern Regressions



Note. (a) A sagittal slice (x = -6) showing the mPFC areas used in the ROI analysis. I defined the mPFC ROI with the term "self-referential" based on Neurosynth term-based meta-analysis. (b) Beta values from the multivariate pattern regression with activation patterns of the self > semantic contrast as a dependent variable. Coloured horizontal lines indicate mean beta values, and lower/upper box limits represent 95% confidence intervals (CIs). (c) Adjusted R^2 from the original regression model (Figure 4.3) and the noise ceiling model. Pink circles indicate mean R^2, and black/grey circles indicate R^2 of individual subjects. (d)

Variance in mPFC ROI activation patterns of the self-reference condition that was explained by activation patterns of the other, introspection, and memory conditions. In panels b and d, bell shaped grey areas indicate permutation distribution.

Adjusted R^2 were significantly lower than the that of the noise ceiling model ($p_{perm} < 0.001$; Figure 4.6c). Hence, there was still unexplained variance even after considering noise in the fMRI data, suggesting that there were patterns of activations specific to the self-reference judgment (not shared by the other three tasks). Activation patterns of the other > semantic, introspection > categorisation, and memory > knowledge contrasts collectively explained, on average, 79.44 % of explainable variances in the mPFC activation patterns of the self > semantic contrast.

4.4.2.4.3 Variance Partitioning Analysis (VPA):

I conducted a variance partitioning analysis to quantify how much variance in the mPFC ROI responses of the self > semantic contrast is explained uniquely by activation patterns of each of the other > semantic, introspection > categorisation, and memory > knowledge contrasts, while considered together with the other two conditions. I present the results in Figure 4.6d. Each of the seven portions significantly explained the variance in neural responses of the self > semantic contrast (all $p_{perm} < 0.001$). The results suggest that the mPFC activation patterns reflect multiple cognitive processes. For example, a significant amount of variances explained by all three contrasts indicate that there were specific patterns of mPFC neural responses that were shared across self-reference, other-reference, introspection, and memory tasks, which likely reflects a cognitive process common for the four tasks. Similarly, a significant amount of variances explained by the other-reference and introspection conditions indicate that there were specific patterns of mPFC neural responses that were shared across self-reference, other-reference, and introspection tasks which likely reflects a cognitive process common for the three tasks, but not the memory task (see below for more discussion).

I ran the same multivariate pattern regression and variance partitioning analyses in other regions related to self-reference (based on the same Neurosynth meta-analysis map with the term "self-referential") and to the default mode network (based on Andrews-Hanna et al., 2010). The anterior and dorsal parts of the mPFC (amPFC and dmPFC) of the default mode network were the only regions that evinced the same pattern of the results as the mPFC reported above: (1) significantly positive beta values for all three independent variables (multivariate pattern regression); (2) significantly positive variance explained for all seven portions (variance partitioning analysis; Figure 4.7), indicating a unique and complex role played by the mPFC during thinking about the self.

Figure 4.7 Results of the Multivariate Pattern Regressions and Variance Partitioning Analyses

	region	MNI coordinate			voxels	Multivariate Regression			Variance Partitioning Analysis						
		x	y	z		Other-ref	Introspection	Memory	1. Other	2. Introspection	3. Memory	4. Other + Introspection	5. Other + memory	6. Introspection + memory	7. All
Self-related regions (Neurosynth)	dmPFC	-6	41	41	56	***	***	***	***	***	***	***	***	*n.s.*	***
	PCC	-3	-58	26	131	***	***	**	***	*n.s.*	***	***	***	***	***
	Left TPJ	-48	-64	29	66	***	***	*n.s.*	***	*n.s.*	*n.s.*	***	***	*n.s.*	***
	Left TP	-60	-13	-22	70	***	***	*n.s.*	***	***	***	***	***	***	*n.s.*
Default mode network (Andrews-Hanna et al., 2010)	vmPFC	0	26	-18	39	***	***	*n.s.*	*	**	*n.s.*	*n.s.*	*	***	*n.s.*
	amPFC	-6	52	-2	93	***	***	***	***	**	***	***	***	**	***
	dmPFC	0	52	26	116	***	***	***	***	***	***	***	***	***	***
	Left TC	-60	-24	-18	117	***	***	*n.s.*	***	**	***	***	***	***	*n.s.*
	Left TPJ	-54	-54	28	110	***	***	*n.s.*	***	*n.s.*	*n.s.*	***	***	*n.s.*	***
	PCC	-8	-56	26	116	***	***	**	***	*n.s.*	***	***	***	***	***
	PHC	-28	-40	-12	79	*n.s.*	***	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
	pIPL	-44	-74	32	78	***	***	*n.s.*	***	*n.s.*	*n.s.*	***	***	*n.s.*	***
	Rsp	-14	-52	8	83	*n.s.*	***	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
	TempP	-50	14	-40	21	*n.s.*	***	***	*	**	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*

Note. I defined the self-related brain regions as "self-referential" based on the Neurosynth meta-analysis map. Regions within the default mode network were based on Andrews-Hanna et al. (2010). For the self-related ROIs, I used all voxels within each cluster. For the ROIs from the default mode network, I used a 9-mm sphere surrounding the center coordinate (maximum of 123 voxels). *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$ (uncorrected). All p values rely on permutation test (1,000 times). *n.s.* non-significant. dmPFC, dorsomedial prefrontal cortex. PCC, posterior cingulate cortex. TPJ, temporoparietal junction. TempP, temporal pole. vmPFC, ventromedial prefrontal cortex. amPFC, anterior-medial prefrontal cortex. TC, temporal cortex. PHC, parahippocampal cortex. pIPL, posterior inferior parietal lobule. Rsp, retrosplenial cortex. TempP, temporal pole.

4.5 Discussion

I provided a more nuanced and precise picture of the mPFC's role during thinking about the self. Replicating prior findings, each of the self-reference, other-reference, introspection, and autographical memory tasks activated the mPFC compared to their corresponding control condition (Figure 4.4). Furthermore, I demonstrated that the relationship between activation patterns during the self-reference task and those of the other three tasks (other-reference, introspection, and autobiographical memory) was intricate. That is, mPFC neural responses during the self-reference task were not simply similar to one task and different from the other two tasks. Instead, the mPFC neural responses during the self-reference task were both similar and distinct at the same time from each of the other-reference, introspection, and autobiographical memory tasks (Figure 4.5). The mPFC was the only region across the whole brain that showed these patterns of results.

Furthermore, the multivariate pattern regression together with the variance partitioning analyses revealed complex relationships of activation patterns of each of the three other tasks to mPFC neural responses, during the self-reference task (Figure 4.6). According to the variance partitioning analyses, not only each of the other-reference, introspection, and memory tasks uniquely explained significant amounts of variances in mPFC neural responses during the self-reference task, but also each pair of these tasks and all three tasks jointly explained significant amounts of variances of the mPFC neural responses during the self-reference task (Figure 4.6d). Hence, it suggests that there are cognitive processes common to thinking about the self and: (1) each of the three tasks; (2) each pair of the three tasks; (3) all three tasks (thus a total of seven cognitive processes; Table 4.1). In addition, adjusted R^2 of the full model (Figure 4.3) were significantly lower than the that of the noise ceiling model (Figure 4.6c), suggesting that there are mPFC neural responses (i.e., a cognitive process) specific to the self-reference task. Overall, our results indicate that there are at least eight cognitive processes (i.e., seven cognitive processes listed in Table 4.1 plus a self-specific process) at play simultaneously when performing the self-reference task, some of which are common across tasks. Our experiment

does not specify what these cognitive processes are (see Table 4.1 for ideas on possible candidate process), leaving this issue open for future research. Nonetheless, as to the self-specific cognitive process, in our prior experiment, I reported that the self-specific activation patterns depend on the importance of the stimuli for self-concept (Levorsen et al., 2023), and so access to this self-concept information stored in the mPFC may be responsible for the self-specific mPFC activation patterns I observed here.

Table 4.1 Seven Cognitive Processes Shared Across Self-Reference and Other Tasks, and Their Possible Candidates

	Cognitive process shared across self-reference and other tasks	**Possible candidates**
1	Cognitive process common to self- and other-reference tasks	• Evaluating consistency with internal goal (i.e., feeling good about oneself by enhancing self and one's friend) • Evaluating personal/social relevance
2	Cognitive process common to self-reference and introspection tasks	• Paying attention to one's internal state (i.e., introspection)
3	Cognitive process common to self-reference and memory tasks	• Retrieval of personal memories
4	Cognitive process common to self-reference, other-reference, and introspection tasks	• Inferential processing
5	Cognitive process common to self-reference, other-reference, and memory tasks	?
6	Cognitive process common to self-reference, introspection, and memory tasks	?

7	Cognitive process common to self-reference, other-reference, introspection, and memory tasks	• Judgment based on internally constructed representations • Emotion regulation

Note. The variance partitioning analysis showed that variances explained by each of the seven portions were all significantly positive (Figure 4.6d), suggesting that there are at least seven different cognitive processes at play simultaneously in the mPFC (plus self-specific process; see Discussion). Note that possible candidates listed here are purely speculative, and I have no intention to claim that these processes are responsible for the result.

This view of the mPFC's role in the self-reference task invites re-interpretation of prior findings. For example, a few experiments showed that mPFC activation patterns are different depending on the target person during the self/other-reference tasks (e.g., self versus close-other versus distant other) and dimensions of person knowledge (e.g., traits, physical attributes, social roles; Courtney & Meyer, 2020; Feng et al., 2018; Koski et al., 2020). The present experiment suggests that what drives these different mPFC neural responses might be differences in how much each task relies on different information (thus, cognitive processes). For instance, thinking about close others and acquaintances might rely more on one's autobiographical memory, whereas thinking about unfamiliar others (e.g., celebrities) might rely more on semantic memory (Courtney & Meyer, 2020). The mPFC activation patterns are also likely to vary depending on whether a context is general or specific (I am friendly in general versus at the university; Martial et al., 2018) and differences in various dimensions of distance similarity (e.g., temporal, spatial, social, hypothetical) (Tamir & Mitchell, 2011) as these judgements likely rely on distinct sources of information.

Moreover, our variance partitioning analysis showed that, among regions related to self-reference and regions in the default mode network, the mPFC is the only region that showed

significantly positive variance explained for all seven portions (Figure 4.6d). This result suggests that the mPFC, one of the core hubs of the default mode network (Andrews-Hanna, 2012; Andrews-Hanna et al., 2010), might be a place where necessary information is gathered and integrated for judgments based on internally constructed representations. As a metaphor, to make a soup, one needs to gather ingredients from different parts of the kitchen and mix them in a pot, with different soups often having some common ingredients (e.g., Italian minestrone and Japanese miso soup commonly use some vegetables, and all soups use water). Similarly, to perform a task that requires a decision based on internally constructed representations (Andrews-Hanna et al., 2014; Buckner & DiNicola, 2019; Menon, 2023; Nakao et al., 2012; Wen et al., 2020), necessary information is gathered from different parts of the brain and integrated in the mPFC. Just like Italian minestrone and Japanese miso soup, different tasks often rely on common cognitive processes (e.g., autobiographical memory and introspection), and there may be a common cognitive process(es) for all such tasks, like water for all soups. This view offers an insight into why diverse social and cognitive tasks activate the mPFC. The mPFC has been consistently implicated not only in the four tasks (self-reference, other-reference, introspection, and autobiographical memory) used in the present experiment, but also in other tasks such as theory of mind, episodic future thinking, and spatial navigation (Andrews-Hanna et al., 2014; Buckner & DiNicola, 2019; Menon, 2023; Wen et al., 2020). It is likely that to perform these tasks one needs to gather information from different part of the brain so as to construct internal representations.

This idea for the role of the mPFC role is largely consistent with roles of the default mode network proposed previously. For example, Yeshurun et al. (2021) considered the default mode network as an active and dynamic sense-making network that integrates incoming extrinsic information with prior intrinsic information to form rich, context-dependent models of situations as they unfold over time. More recently, Menon (2023) argued that the default mode network integrates multiple cognitive functions to create a coherent internal narrative of our experiences. Within these large frameworks on the default mode network function (see also

Koban et al., 2021), the present experiment provides evidence that, among regions in the default mode network, the mPFC is a place where all information converges and is integrated to form coherent internal representation for making a task-relevant judgement. Put otherwise, multiple cognitive processes are performed in the mPFC during a single task (i.e., self-reference task).

Our results highlight an important conceptual challenge for social/cognitive neuroscientists; each of many tasks used in the field involves multiple cognitive processes (or operations), and each of these processes needs to be identified to fully understand the function of the mPFC (and any other brain regions). For example, our findings indicate that the difference between the self-reference and semantic tasks is not only the level of self-referential processing, but also that there are several other additional cognitive processes involved in the self-reference task, some of which are shared with other-reference, introspection, and memory tasks (see Table 4.1). Thus, instead of a traditional brain mapping approach (showing which regions are activated by the self-reference > semantic contrast), what is required is brain mapping at a much finer scale; in which more basic cognitive process is linked to specific activation patterns within an area. Although the utility of a multivariate approach over a univariate approach has been well recognised (and its methodology has been well developed) (Haxby, 2012; Haynes & Rees, 2006), identifying each of various basic cognitive processes involved in a social/cognitive task remains a challenge (Schaafsma et al., 2015). The multivariate pattern regression approach (with the variance partitioning analysis) where both dependent and independent variables are neural responses (Figure 4.3) may be a good approach as it helps us at least to statistically decompose complex social/cognitive tasks and identify whether there are unique or common processes across different tasks.

In conclusion, the current findings enhance understanding of the mPFC and its involvement in self-referential thinking by demonstrating its unique role in integrating diverse cognitive processes. The mPFC is not merely activated by self-reference, but also shows complex activation patterns that are both similar and distinct from other cognitive tasks such as other-reference, introspection, and autobiographical memory. Taken together with the role of

the mPFC within the default mode network reported previously, the findings indicate that the mPFC serves as a hub where information from various brain regions is gathered and integrated, facilitating tasks that involve constructing internal representations.

Chapter 5 General Discussion

5.1 Aims

The primary objective of the thesis was to enhance understanding of how the self is represented in the brain. In this chapter, I will summarise the main findings and discuss the contribution of the thesis to the literature. Next, I will consider the strengths and limitations of this thesis along with future directions for neuroimaging research about the self. Finally, I add a note on reproducibility, before concluding.

5.2 Summary of Key Findings

In Chapter 2, I presented two fMRI experiments that addressed how the self is represented in the brain. In Experiment 1, participants carried out a self-reference task and a word-class judgement task. By using RSA, I examined how self-importance and self-descriptiveness are represented in the mPFC, while controlling for potential confounding factors. I found that self-importance, but not self-descriptiveness, is represented in the mPFC. To test if this finding is self-specific, I conducted Experiment 2 where participants also carried out an other-reference task, where "other" was a close other. I found that the importance of a stimulus for our own identity, but not that of a close other, is represented in the mPFC. Overall, the findings of Chapter 2 suggest that the self-concept is represented in the mPFC in terms of self-importance.

In Chapter 3, I reported a behavioural reaction time experiment, addressing how the self-concept is represented in an associative memory structure. Based on the Chapter 2 findings, I tested whether the associative links in an associative memory structure of the self are represented in terms of self-importance. I used an evaluative priming paradigm. I hypothesised that self-importance, but not self-descriptiveness, would show a priming facilitation effect for self-related stimuli compared to other-related stimuli. The results did not support the

hypothesis. Exploratory analyses revealed a facilitation effect of self-descriptiveness on reaction time, suggesting that links in a memory structure are represented in terms of self-descriptiveness.

In Chapter 4, I reported an fMRI experiment that examined the extent to which patterns of activation in the mPFC during the self-reference task can be explained by patterns of activation during other social or cognitive tasks such as other-reference, introspection, and autobiographical memory. I discovered that there were both patterns of activation that are shared for self-reference and each of the three other tasks, but also patterns of activation that are specific to the self. These results suggest that, during self-reference, mPFC is recruited for processes shared with other tasks. The results are also consistent with the idea that mPFC works as a hub in the DMN. In the next section, I outline how the findings contribute to the literature on the neural representation of the self.

5.3 Contribution of the PhD Thesis

5.3.1 Is the Self Special in the Brain?

As detailed in Chapter 1, a major question in neuroimaging research about the self has been whether the self is special in the brain (Kelley et al., 2002; Macrae et al., 2004; Ochsner et al., 2005). The question was posed by early psychological research on the self (Rogers et al., 1977) and was re-posed by fMRI researchers. These researchers (e.g., Kelley et al., 2002; Macrae et al., 2004) argued that, because activation during the self-reference task is greater than activation during control conditions in the mPFC, a region separate from those regions previously known to process semantic information, self-reference processing is distinguishable from other processes in the brain . However, the idea that the self is specifically processed in the mPFC was first questioned by experiments that did not replicate the findings in the close-other (Schmitz et al., 2004) or similar-other (Krienen et al., 2010) conditions. Furthermore, in a literature review, Legrand and Ruby (2009) argued that mPFC activation during the self-reference task is not related to the self specifically, but is merely a result of general cognitive

processes taking place during the self-reference task, as well as during several other tasks (e.g., reward, autobiographical memory).

The findings of Chapters 2 and 4 are pertinent to the issue of self-specificity in the mPFC. In Chapter 2, I asked what information about the self that is represented in the mPFC and whether this representation is self-specific. To find out, I compared self-reference judgements to judgements of a close friend. The results enriched the literature by demonstrating that the mPFC represents the importance of a stimulus to the self-concept and that this representation is self-specific. In Chapter 4, I compared the self-reference task to an other-reference, autobiographical memory, and introspection task. While there were some patterns of activation shared for self-reference and each of the other tasks in the mPFC, there were also patterns of activation in the mPFC that were specific to the self-reference task. Since we observed both self-specific activation patterns in the mPFC and activation patterns shared across tasks, our findings overall support both sides of the abovementioned debate. (Kelley et al., 2002; Macrae et al., 2004; Ochsner et al., 2005; Legrand & Ruby, 2009).

Specifically, our findings show self-specific activation patterns in the mPFC, consistent with early claims that the self is special in the brain (Rogers et al., 1977) and processed in the mPFC (Kelley et al., 2002; Macrae et al., 2004; Ochsner et al., 2005). At the same time, because we also observed activation patterns in the mPFC shared across tasks, our results align with arguments that mPFC activation during the self-reference task reflects general cognitive processes (Legrand & Ruby, 2009). Thus, rather than supporting one side of the debate, our results suggest a more complex mPFC function that integrates both perspectives.

In all, the thesis makes three contributions. First, Chapters 2 and 4 suggest that self-related information is specifically processed in the mPFC (compared to an other-reference task in Chapter 2, and an other-reference task, autobiographical memory task, and introspection task, in Chapter 4). These results are consistent with a neuropsychological lesion study (Marquine et al., 2016), suggesting a self-specific role of the mPFC; in particular, the mPFC was

necessary for processing self-reference, but not other-reference of a close other. The results are also consistent with previous multivariate experiments indicating that the self-reference and other-reference tasks are represented by distinguishable patterns of activation in the mPFC (Courtney & Meyer, 2020; Feng et al., 2018; Koski et al., 2020; Parelman et al., 2022). Chapters 2 and 4 extend these findings as we also find that when compared to autobiographical memory and introspection, there are patterns of activation that are self-specific in the mPFC.

Second, the results deepen the understanding of what information about the self that is represented in the self-specific activation in the mPFC. As extensively reviewed in Chapter 1, given that the mPFC is consistently active during the self-reference task (Wagner et al., 2012;2019), neuroimaging researchers have long questioned what information about the self that is represented in the mPFC during the task. For example, researchers have proposed that mPFC activation during the self-reference task might be explained for by the self-relevance of the stimuli presented (Phan et al., 2004). Others have proposed that it is the reward (Enzi et al., 2009) or value D'Argembeau et al. (2012) of the stimuli that is processed during the self-reference task. Furthermore, researchers have proposed that autobiographical memory (Araujo et al., 2014), and introspection (Goldberg et al., 2006) explain the mPFC activation during the self-reference task. The findings presented in Chapter 2 contribute to the literature by answering this long-standing question. Our findings demonstrate that it is how important a stimulus is to one's identity that is represented in the mPFC during the self-reference task, and that this representation is self-specific. In other words, in the extensive research reviewed in Chapter 1- where participants judged the self-descriptiveness of stimuli-our findings suggest that what was processed in the mPFC in these experiments was the self-importance of the stimuli.

Third, the results align with the idea that there are general cognitive processes taking place during the self-reference task that are shared across other tasks, which result in mPFC activation (Legrand & Ruby, 2009). Chapter 4 extends previous research by using MVPA to compare the self-reference task to other tasks which had previously only been compared with univariate analysis (i.e., autobiographical memory, introspection). The findings suggest that

there are shared patterns of activation (cognitive processes) in the mPFC when self-reference is compared to other-reference, autobiographical memory and introspection.

Overall, our results in Chapters 2 and 4 contribute to the literature with a nuanced understanding of mPFC activation during the self-reference task. When comparing the self-reference task to other-reference, autobiographical memory, and introspection, patterns of activation are both similar and different. As such, mPFC activation during the self-reference task represents both shared cognitive processes across tasks and patterns of activation that are self-specific. Furthermore, the self-specific patterns of activation represent self-importance.

5.3.2 What Is the Role of the mPFC in the Default Mode Network?

Previous research suggests that the mPFC is active during various tasks (amongst others; self-reference, other-reference, autobiographical memory and introspection) (Kelley et al., 2002; Denny et al., 2012; Martinelli et al., 2013; Goldberg et al., 2006). Research has also proposed that the mPFC functions like a central hub in the DMN (Andrews-Hanna et al., 2010). Our findings presented in Chapter 4 demonstrate that mPFC is the only region where patterns of activation are simultaneously both shared and unique across tasks. Specifically, we found that there are at least 7 cognitive processes in the mPFC that are shared across self-reference and other tasks (other-reference, autobiographical memory and introspection). These findings are consistent with the idea that the mPFC works as a central hub in the DMN (Andrews-Hanna et al., 2010), extending the understanding of the mPFC's functions in the DMN.

Taken together, the research in the current thesis contributes to understanding the neural representation of the self by suggesting that the self is represented in the mPFC both by shared cognitive processes and self-specific patterns of activation. Where the shared cognitive processes are involved in the DMN's function as a central hub. Whereas the self-specific patterns of activation in the mPFC process the self-importance of stimuli. The following section will discuss the strengths, limitations, and future directions of this research.

5.3.3 Cultural Context

Previous research (Zhang et al., 2006; Zhu et al., 2007) has proposed that the neural basis of self-representation is influenced by culture. For example, Zhu et al. (2007) reported greater mPFC activation when participants from a Western sample judged adjectives about themselves compared to a close other (their mother), whereas no such difference was observed in an East Asian sample. These findings were interpreted as suggesting mPFC activation is self-specific in Western participants, but reflects both self and mother in East Asian participants. However, as outlined in Chapter 1, the broader literature of univariate experiments on self–other activation overlap in the mPFC has been mixed overall (Schmitz et al. (2004); Ochsner et al. (2005); Heatherton et al., 2006; Chen et al., 2013). Moreover, Zhu et al. (2007) relied on univariate analysis, which is limited in its ability to capture shared neural mechanisms across tasks (Woo et al., 2014), that is, it cannot determine whether self- and mother-related representations are based on a common neural mechanism in the East Asian sample.

In contrast, the experiments included in the present thesis employed multivariate analyses, which provide greater sensitivity in identifying common representational patterns (Woo et al., 2014). Across two experiments (Experiment 2 in Chapter 2 and the experiment in Chapter 3), we found evidence of self-specific activation in the medial prefrontal cortex (mPFC) within an East Asian sample. Taken together, the use of more powerful analytic methods and the consistent observation of self-specific activation in East Asian participants suggest that our findings are unlikely to be affected for by cultural differences.

5.4 Strengths, Limitations, and Future Directions

An obvious strength of the neuroimaging experiments (Chapters 2 and 4) is the use of MVPA. Although MVPA has increasingly been employed to investigate the neural representation of the self (Wagner et al., 2019), it is still worth pointing out the advantages of this approach. These include asking a wider range of questions, rigorously controlling for potential confounds,

and providing more sensitive information about underlying patterns of activation (Dimsdale-Zucker & Ranganath, 2019; Kriegeskorte et al., 2008; Norman et al., 2006).

Another strength of the thesis is its approach to capturing each participants' self-concepts. Instead of relying on a standardised set of personality traits, participants provided their own self-related items, ensuing a broader and more individualised representation. In Chapters 2 and 3, I incorporated self-provided stimuli into the self-reference tasks and the evaluative priming task, respectively. This approach facilitated the inclusion of stimuli with a broad spectrum of ratings concerning self-importance and self-descriptiveness.

Furthermore, in Chapters 2 and 3, I dissociated the effects of the independent variables. Specifically, I created stimuli sets with the goal of achieving the lowest possible intercorrelation in participants' ratings. To achieve this, for each participant, I first selected 40 out of the 80 items they had rated on self-importance, self-descriptiveness, autobiographical memory, valence, and familiarity in Chapter 2, and on self-importance, self-descriptiveness and valence in Chapter 3. I also added whether the item was self-provided or not (Chapters 2 and 3) and the number of characters (Chapter 2). For the 40 items, I calculated correlations across the ratings and recorded the highest correlation coefficient. After repeating this procedure up to 1,000,000,000 times, I selected the 40 items with the lowest correlations. In this way, I managed to dissociate self-importance from self-descriptiveness as well as from the other potentially confounding factors.

Similarly, in the autobiographical memory task of Chapter 4, participants viewed, while in the scanner, previously provided personal memories. Relevant research has typically used a cued recall task in which participants are presented with a word or image to cue the engagement of autobiographical memory processing (Hughes et al., 2024). Compared to this, the task I used in Chapter 4 is more likely to evoke personal memories. In a cued recall task, participants may struggle to retrieve memories associated with certain stimuli. This difficulty can lead them to generate information rather than accessing the actual memory, potentially

resulting in inaccuracies. The phenomenon is related to cue-dependent forgetting, where the absence of effective retrieval cues hinders memory recall. Incorporating self-generated memories ensures that stimuli are personally relevant to participants, thereby increasing the likelihood of authentic memory processing during scanning sessions. By using individualised stimuli, this thesis improved the assessment of self-concept and autobiographical memory.

In addition, with the exception of the first experiment reported in Chapter 2, all experiments were preregistered. This practice enhanced the credibility of the research by differentiating between exploratory and confirmatory analyses. Furthermore, the accompanying transparency enables subsequent researchers to engage in exact replications (Van't Veer & Giner-Sorolla, 2016).

The thesis also has limitations. I tested only healthy student participants within a narrow age range. Although the self-concept is relatively stable across time (Diehl et al., 2006), it may change more rapidly at certain life stages like adolescence (Van Buuren et al., 2022) and in clinical groups (Huang et al., 2017). Follow-up work could address the neural representation of the self in adolescents, older adults, and clinical samples.

For example, future research could examine the neural representation of self-concept also in younger samples. Based on our findings in Chapter 2 which suggest that the self is specifically represented in the mPFC in terms of self-importance in healthy adults, future experimental paradigms could investigate if our results also replicate in samples of adolescents and children. Such experiments could shed light on whether the mPFC specifically represents the self also during early development, a period when close others exert a stronger influence on one's life (Popov et al., 2015).

Further, Chapter 3 did not provide clear answers as to how the self-concept is represented in memory. Previous research has used EEG to investigate the timing (event related potentials) associated with processing self-related information, including valence (Huang et al.,

2025). Future investigations could use EEG to understand potential differences in the timing of the neural activation for processing self-importance, self-descriptiveness, and valence.

It is also worth zeroing in on the functions of the DMN. In his 2023 review, Menon highlighted that research on various concepts related to the DMN—such as mind-wandering, social cognition, and self-referential processing—has often been conducted in isolation. Synthesising this literature, he proposes a model of the DMN where its function is to create an inner narrative and a sense of self. Future work would do well to test this model.

As mentioned in Chapter 1, Meyer and Lieberman (2018) proposed that activation in the mPFC at rest primed thinking about the self and others. The authors argued that priming occurred by default, because people are social, so that at rest the mPFC becomes active which nudges people to think about themselves and others. DMN activation at rest is related to consolidation of social information (Meyer et al., 2019), and social (compared to non-social) information is prioritised in memory consolidation in the dmPFC (Jimenez & Meyer, 2024). These findings point to interesting directions, such as testing individual difference in DMN social memory consolidation and processes related to the self.

Iyer et al. (2024) examined individual differences in DMN activation during rest after viewing negative stimuli. Participants watched videos of terminally ill patients discussing their illness, then recalled and listed what they remembered. Recall was later coded as positive or negative. Participants who remembered the positives showed distinct DMN connectivity during rest, whereas participants who recalled the negatives showed similar DMN connectivity. Hence, people who see the good in the bad during processing of negative social stimuli engage the DMN idiosyncratically during rest, but not during encoding.

Due to reliability and validity concerns, fMRI research has mainly focused on group-level effects rather than individual differences. In individual differences measures of fMRI data, reliability is reduced by noise such as variability in head motion, physiology and vascular substances across participants. And a specific validity challenge is ensuring that the regions

compared across subjects are functionally homologous (Dubois & Adolphs, 2016). Recently there have been advances in analysis tools to measure individual differences in fMRI data, which reduce the abovementioned issues. One such tool is inter-subject representational analysis (IS-RSA) (Chen et al., 2020; Chen & Qu, 2021; Finn et al., 2020), as used in the abovementioned study by Iyer et al. (2024). IS-RSA is conducted by first calculating an intersubject similarity matrix for each type of measurement, such as neural activity and self-reported questionnaires. Next, the correlation between the two similarity matrices is calculated to identify brain regions where participants with similar questionnaire responses also show similar neural patterns. Psychologists have long recognized that strong theories must explain not only the average patterns of behaviour in a representative population, but also the variations that exist across individuals (Underwood, 1975). The development of analysis tools such as IS-RSA enhances neuroscientists' ability to capture individual differences in brain-behaviour relationships (Finn et al., 2020).

IS-RSA can also be used to test hypotheses related to the neural representation of the self. Previous psychological research found that people self-enhance (i.e., perceive the self in an inflated manner) and self-protect (i.e., minimize the negativity of the self; Alicke & Sedikides, 2009; Sedikides & Gregg, 2008). They also engage in various strategies to view ambiguous information in a positive light (Hepper & Sedikides, 2012; Sedikides, 2020). Also, after receiving negative feedback, they reconstruct the information to feel less negative over time (Skowronski et al., 2014). Failure to engage in such processes is associated with psychological maladjustment (Dufner et al., 2019).

Similarly to the paradigm in Iyer et al.'s experiment, one could combine IS-RSA with measures of self-esteem to test the hypothesis that, when processing negative feedback, individuals with high self-esteem take longer to consolidate negative self-related information. During subsequent rest, they may show idiosyncratic DMN connectivity, possibly reflecting the engagement of self-enhancement strategies that reframe negative information in a more positive light.

Kanske et al. (2017) reported that greater mind-wandering is positively related to narcissism, with high (compared to low) narcissists mind-wandering about self-indulgent success to a greater extent. Future research could test how individual differences in mind-wandering or narcissism are associated with neural processes at rest.

The idea that we are social by default and think about ourselves and other people readily at rest (Lieberman & Meyer, 2018; Meyer, 2019; Jimenez & Meyer, 2024) can also create research question in relation to self-reference memory processes. As outlined in Chapter 1, the behavioural literature on the self-reference effect in memory is extensive (Symons & Johnson, 1997), and a few neuroimaging experiments also examined it (Macrae et al., 2004; Kim & Johnson, 2012; Koski et al., 2020). A possible explanation for the self-reference memory effect is that since we so quickly default to thinking about ourselves and consolidate social information immediately at rest (Meyer & Lieberman, 2018), perhaps the self-reference effect occurs because we can so readily integrate new information about the self as we are already consolidating information about it.

In the 1980s, with the increasing use of computers in experiments, social psychologists started to borrow methods from cognitive science and psychophysics. For example, they presented participants with social (versus non-social) stimuli on a computer screen. The advent of fMRI facilitated the development of social cognitive neuroscience. However, using fMRI to investigate social interactions has been challenging, as participants are tested in isolation inside a noisy scanner allowing for minimal movement. Recently, however, social cognitive neuroscience has shifted focus from examining participants in isolation to paradigms that allows for the measure of interacting minds (Wheatley, 2024). One approach to studying the neural mechanisms of social interactions has been fMRI hyperscanning. Here, neural activation is measured from multiple participants simultaneously (Misaki et al., 2021). This approach investigates neural activation of a social interactions in real-time. Two-way real interactions and shared experiences evoke neural processes that cannot be captured by scanning a person in isolation

(Tsoi, 2022). Experiments using this approach can address such questions as whether mental representations during a conversation are related to social outcomes (e.g., liking).

Likewise, researchers could employ neuroimaging methods in combination with social network analysis. Parkinson et al. (2018) used fMRI to scan participants while they were watching a movie. Participants were part of a shared real-world social network. They also completed questionnaires to quantify their social network. Parkinson et al. found that participants who were closer showed more similar neural activation while watching the video, capturing how neural activation is related to real-world-friendships.

Overall, neuroimaging research has recently used paradigms and methods that allow for measures of social interactions. Experiments have been increasingly conducted in such settings as dyads (Zhang et al., 2021) and groups (Guthrie et al., 2022). Further, neuroimaging research on social cognition has increasingly employed functional Near Infrared-Spectroscopy (fNIRS), a neuroimaging technique that is portable and wireless, thus easily used outside the laboratory and in everyday social settings (Pinti et al., 2020). In all, neuroimaging has advanced to be more ecologically valid and generalisable to settings outside the scanner.

Neuroimaging research on social cognition has become increasingly social under the understanding that focusing on a single organism is insufficient to fully appreciate the biological processes in social species (Wheatley et al., 2024). Thus, with the rapid progress in techniques for examining brain activity during social interactions, future research in social cognitive neuroscience should continue to examine brains in social contexts.

5.5 A Note on Reproducibility in Neuroimaging Research

In the past fifteen years, there has been an increased focus on reproducibility in science (Baker, 2016), including neuroimaging research (Szucs & Ioannidis, 2020). Although the sample sizes in fMRI experiments have gradually increased over the years (by 0.74 participants each year) (Szucs & Ioannidis, 2020), they are still relatively low (Poldrack et al., 2017). Low statistical

power not only decreases the chances of detecting a real effect when it exists, but also increases the probability that any detected positive result is actually false, while at the same time inflating the size of observed effects (Yarkoni, 2009). Although power calculations are important, they appeared in only 4% of fMRI experiments published in 2017 and 2018 (Szucs & Ioannidis, 2020).

Another important practice for improving reproducibility in neuroimaging research is preregistration. This practice enhanced the credibility of the research by differentiating between exploratory and confirmatory analyses. Furthermore, the accompanying transparency enables subsequent researchers to engage in exact replications (Van't Veer & Giner-Sorolla, 2016). In addition to preregistrations and power calculations, practices that enhance reproducibility, transparency, and impact include sharing data, code, preprints, and all supplementary analyses (Gorgolewski & Poldrack, 2016).

Of the three fMRI experiments included in this thesis, one experiment included a power analysis and two experiments were preregistered. In addition, all three experiments have data available in repositories and all analyses were either included in the main papers or supplementary files. Furthermore, preprints were made publicly available, prior to peer-reviewed publications. As reproducibility is essential for scientific progress (Turner et al., 2018), it is critical that future neuroimaging research continues to implement practices that enhance transparency (Poldrack et al., 2017).

5.6 Conclusion

The findings presented suggest that the self is specifically represented in the mPFC in terms of self-importance. Moreover, it also suggests that general cognitive processes take place during the self-reference task, which are shared with other tasks such as other-reference, autobiographical memory and introspection. This latter finding is consistent with the function of the mPFC as a central hub in the DMN. The thesis deepened understanding of the self in the brain.

Appendix A Supplementary Material for Chapter 2 Participants Instructions in the first questionnaire

Supplementary texts

Instruction given at the first online questionnaire: At the beginning of the first online questionnaire, participants were given the following instruction (translated from the original texts written in Japanese by DeepL [https://www.deepl.com/en/translator]).

In this first questionnaire, we will ask you who you are.

Please feel free to answer in the format "I _____."

Please feel free to write about anything that applies to you, such as your physical characteristics (e.g., I am tall), your personality (e.g., I am social), what you like (food, music, artists, etc.; e.g., I like xx) or dislike (I dislike xx), groups you belong to (university, department, clubs, etc.; e.g., I belong to xx clubs), your background (e.g., I graduated from xx High School), values/ideas that you hold dear (e.g., I am an animal rights activist), and so on.

Please tell us not only about your likes and strengths, but also about your dislikes, weaknesses, complexes you have, and other negative things.

You may start a sentence with "I" or "My" (e.g., my favorite word is "xx," "my birthday is on March xx," "my dog's name is xx," etc.).

However, please do not include any information that completely identifies you (e.g., name, e-mail address, etc.). If you cannot be completely identified, please be as specific as possible (e.g., name of high school, clubs you belong to, etc.).

Your answers will be associated with a random ID number and will be stored separately from your name, e-mail address, and other personally identifiable data. However, you do not have to give us anything that you do not want others to know. Please only describe things that you are willing to share.

Some of the words and phrases that you give will be used in a second online questionnaire and fMRI experiment at a later date. The information you provide here will never be used for any purpose other than the experiment.

Please give a minimum of 30 phrases (maximum 40).

When you are ready, please fill in and say a sentence starting with "I" in the answer box below.

Appendix B Supplementary Material for Chapter 3

Participants Instructions in the first questionnaire

Please enter your participant ID number below

In this first questionnaire, we will ask you who you are.
Please answer in the format "I _____."

Please feel free to write about anything that applies to you, such as your physical characteristics (e.g., I am tall), your personality (e.g., I am social), what you like (food, music, artists, etc.; e.g., I like xx) or dislike (I dislike xx), groups you belong to (university, department, clubs, etc.; e.g., I belong to xx clubs), your background (e.g., I graduated from xx High School), values/ideas that you hold dear (e.g., I am an animal rights activist), and so on.

Please tell us not only about your likes and strengths, but also about your dislikes, weaknesses, complexes you have, and other negative things.

You may start a sentence with "I" or "My" (e.g., my favourite word is "xx," "my birthday is on March xx," "my dog's name is xx," etc.). However, please do not include any information that completely identifies you (e.g., name, e-mail address, etc.). Although you should not include information that completely identifies you, please try to be as specific as possible (e.g., name of high school, clubs you belong to, etc.).

Your answers will be associated with your (random) participant ID number and will be stored separately from your name, e-mail address, and other personally identifiable data. However, you do not have to give us anything that you do not want others to know. Please only describe things that you are willing to share.

Some of the words and phrases that you give will be used in a second online questionnaire and the behavioural experiment at a later date. The information you provide here will not be shared with anyone outside the research team and will never be used for any purpose other than the experiment.

Please provide a minimum of 30 phrases (maximum 40). When you are ready, please click the button below to start to fill in.

Please provide sentences with information about yourself in the answer boxes below.

Please feel free to write about anything that applies to you, for example:

- physical characteristics (e.g., I am tall)
- your personality (e.g., I am social)
- what you like (food, music, artists, etc.; e.g., I like xx)
- what you dislike (I dislike xx)
- groups you belong to (university, department, clubs, etc.; e.g., I belong to xx clubs)
- your background (e.g., I graduated from xx High School)
- values/ideas that you hold dear (e.g., I am an animal rights activist), and so on

Please write in the format "I _____." Please provide at least 30 sentences.

1)

2)

Appendix C Supplementary Material for Chapter 3 Post-Experimental Questionnaire

1. GENDER: _____

2. AGE: _____

3. RACE/ETHNICITY: (check as many general categories that apply and specify all of possible details)

AFRICAN_____

ASIAN_____

CAUCASIAN_____

HISPANIC/LATINO_____

INDIAN (India)_____

MIDDLE EASTERN_____

SOUTH AMERICAN_____

OTHER_____

4. PLACE PRIMARILY RAISED: (City & Country)

_____

5. NUMBER OF YEARS YOU HAVE BEEN A STUDENT AT THE UNIVERSITY OF

SOUTHAMPTON: _____

6. ENTER YOUR NATIVE LANGUAGE: _____

7. DO YOU CONSIDER YOURSELF TO BE:

HETEROSEXUAL OR STRAIGHT _____

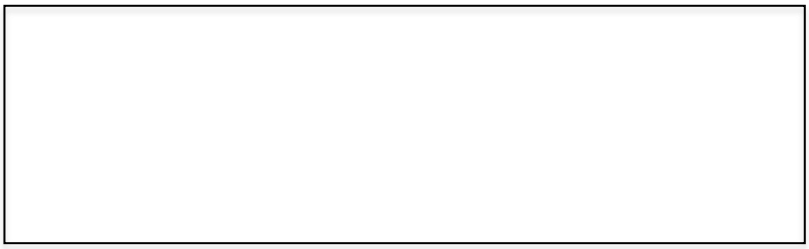
GAY OR LESBIAN _____

BISEXUAL_____

8. HOW MUCH DID YOU SLEEP LAST NIGHT?: _____hours

9. HANDEDNESS: RIGHT_____ LEFT_____

10. What do you think is the purpose of this study? (optional)



Thank you

List of References

Addis, D. R., Wong, A. T., & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, *45*(7), 1363–1377. https://doi.org/10.1016/j.neuropsychologia.2006.10.016

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, *20*(1), 1–48. https://doi.org/10.1080/10463280802613866

Alink, A., Walther, A., Krugliak, A., van den Bosch, J. J. F., & Kriegeskorte, N. (2015). Mind the drift - improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. *BioRxiv*, 032391. http://biorxiv.org/content/early/2015/11/20/032391.abstract

Alves, P. N., Foulon, C., Karolis, V., Bzdok, D., Margulies, D. S., Volle, E., & Thiebaut de Schotten, M. (2019). An improved neuroanatomical model of the default-mode network reconciles previous neuroimaging and neuropathological findings. *Communications Biology*, *2*(1), 1–14. https://doi.org/10.1038/s42003-019-0611-3

Andersen, S. M., Reznik, I., & Manzella, L. M. (1996). Eliciting facial affect, motivation, and expectancies in transference: Significant-other representations in social relations. *Journal of Personality and Social Psychology, 71*(6), 1108–1129. https://doi.org/10.1037/0022-3514.71.6.1108

Anderson, J.R. (1976). *Language, Memory and Thought*. Erlbaum.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, *22*(3), 261-295. https://doi.org/10.1016/S0022-5371(83)90201-3

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Winston.

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology, 9*(3), 272–279. https://doi.org/10.1037/h0025907

Anderson, N.H. (1974). Cognitive algebra: Integration theory applied to social attribution. *Advances in Experimental Social Psychology, 7*, 1–101. https://doi.org/10.1016/S0065-2601(08)60035-0

Anderson, N. H. (1981). *Foundations of information integration theory*. Academic Press.

Anderson, S. J., & Conway, M. A. (1993). Investigating the structure of autobiographical memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(5), 1178–1196. https://doi.org/10.1037/0278-7393.19.5.1178

Andrews-Hanna, J. R. (2012). The brain's default network and its adaptive role in internal mentation. *Neuroscientist*, *18*(3), 251–270. https://doi.org/10.1177/1073858411403316

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., & Buckner, R. L. (2010). Functional-Anatomic Fractionation of the Brain's Default Network. *Neuron*, *65*(4), 550–562. https://doi.org/10.1016/j.neuron.2010.02.005

Andrews-Hanna, J. R., Smallwood, J., & Spreng, R. N. (2014). The default network and self-generated thought: Component processes, dynamic control, and clinical relevance. *Annals of the New York Academy of Sciences*, *1316*(1), 29–52. https://doi.org/10.1111/nyas.12360

Araujo, H. F., Kaplan, J., Damasio, H., & Damasio, A. (2014). Involvement of cortical midline structures in the processing of autobiographical information. *PeerJ*, *2014*(1), 1–26. https://doi.org/10.7717/peerj.481

Araujo, H. F., Kaplan, J., Damasio, H., & Damasio, A. (2015). Neural correlates of different self domains. *Brain and Behavior*, *5*(12), 1–5. https://doi.org/10.1002/brb3.409

Aron, A., McLaughlin-Volpe, T., Mashek, D., Lewandowski, G., Wright, S. C., & Aron, E. N. (2004). Including others in the self. *European Review of Social Psychology*, *15*(1), 101–132. https://doi.org/10.1080/10463280440000008

Baggini J (2011). *The ego trick: What does it mean to be you?* Granta Publications.

Banse, R. (1999). Automatic evaluation of self and significant others: Affective priming in close relationships. *Journal of social and personal relationships*, *16*(6), 803-821. https://doi.org/10.1177/0265407599166007

Banse, R. (2001). Affective priming with liked and disliked persons: Prime visibility determines congruency and incongruency effects. *Cognition and Emotion*, *15*(4), 501–520. https://doi.org/10.1080/02699930126251

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063

Begg, I. (1978). Imagery and organization in memory: Instructional effects. *Memory & Cognition*, *6*(2), 174–183. https://doi.org/10.3758/BF03197443

Benoit, R. G., Gilbert, S. J., Volle, E., & Burgess, P. W. (2010). When I think about me and simulate you: Medial rostral prefrontal cortex and self-referential processes. *NeuroImage*, *50*(3), 1340–1349. https://doi.org/10.1016/j.neuroimage.2009.12.091

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Binning, K. R., Unzueta, M. M., Huo, Y. J., & Molina, L. E. (2009). The interpretation of multiracial status and its relation to social engagement and psychological well-being. *Journal of Social Issues*, *65*(1), 35–49. https://doi.org/10.1111/j.1540-4560.2008.01586.x

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, *79*(4), 631–643. https://doi.org/10.1037/0022-3514.79.4.631

Bower, H., & Gilligan, S. G. (1979). Information Related to One' s Self. *Journal of Research in Personality*, *13*(4), 420–432. http://www.sciencedirect.com/science/article/pii/0092656679900059

Bracci, S., Daniels, N., & De Beeck, H. O. (2017). Task context overrules object- And category-related representational content in the human parietal cortex. *Cerebral Cortex*, *27*(1), 310–321. https://doi.org/10.1093/cercor/bhw419

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*(2), 49–57. https://doi.org/10.1016/j.tics.2006.11.004

Buckner, R. L., & DiNicola, L. M. (2019). The brain's default network: updated anatomy, physiology and evolving insights. *Nature Reviews Neuroscience*, *20*(10), 593–608. https://doi.org/10.1038/s41583-019-0212-7

Burt, J. B., Helmer, M., Shinn, M., Anticevic, A., & Murray, J. D. (2020). Generative modeling of brain maps with spatial autocorrelation. *NeuroImage*, *220*(June), Article 117038. https://doi.org/10.1016/j.neuroimage.2020.117038

Cai, H., Wu, L., Shi, Y., Gu, R., & Sedikides, C. (2016). Self-enhancement among Westerners and Easterners: A cultural neuroscience approach. *Social Cognitive and Affective Neuroscience*, *11*(10), 1569–1578. https://doi.org/10.1093/scan/nsw072

Chavez, R. S., Heatherton, T. F., & Wagner, D. D. (2017). Neural Population Decoding Reveals the Intrinsic Positivity of the Self. *Cerebral Cortex*, *27*(11), 5222–5229. https://doi.org/10.1093/cercor/bhw302

Chen, P. H. A., Jolly, E., Cheong, J. H., & Chang, L. J. (2020). Intersubject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. *NeuroImage*, *216*. https://doi.org/10.1016/j.neuroimage.2020.116851

Chen, P. H. A., & Qu, Y. (2021). Taking a Computational Cultural Neuroscience Approach to Study Parent-Child Similarities in Diverse Cultural Contexts. In *Frontiers in Human Neuroscience* (Vol. 15). Frontiers Media S.A. https://doi.org/10.3389/fnhum.2021.703999

Chen, P.-H. A., Wagner, D. D., Kelley, W. M., Powers, K. E., & Heatherton, T. F. (2013). Medial prefrontal cortex differentiates self from mother in Chinese: evidence from self-motivated immigrants. *Culture and Brain*, *1*(1), 3–15. https://doi.org/10.1007/s40167-013-0001-5

Conneely, M., McNamee, P., Gupta, V., Richardson, J., Priebe, S., Jones, J. M., & Giacco, D. (2021). Understanding Identity Changes in Psychosis: A Systematic Review and Narrative Synthesis. *Schizophrenia Bulletin*, *47*(2), 309–322. https://doi.org/10.1093/schbul/sbaa124

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language*, *53*(4), 594–628. https://doi.org/10.1016/j.jml.2005.08.005

Cooley, C.H. (1902). *Human nature and the social order*. Charles Scribner's Sons.

Cooper, N., Tompson, S., O'Donnell, M. B., & Falk, E. B. (2015). Brain activity in self- and value-related regions in response to online antismoking messages predicts behavior change. *Journal of Media Psychology*, *27*(3), 93–108. https://doi.org/10.1027/1864-1105/a000146

Courtney, A. L., & Meyer, M. L. (2020). Self-Other representation in the social brain reflects social connection. *Journal of Neuroscience*, *40*(29), 5616–5627. https://doi.org/10.1523/JNEUROSCI.2826-19.2020

Cousins, S. D. (1989). Culture and Self-Perception in Japan and the United States. *Journal of Personality and Social Psychology*, *56*(1), 124–131. https://doi.org/10.1037/0022-3514.56.1.124

Craik, F. I., Moroz, T. M., Moscovitch, M., Stuss, D. T., Winocur, G., Tulving, E., & Kapur, S. (1999). In search of the self: A positron emission tomography study. *Psychological Science*, *10*(1), 26-34. https://doi.org/10.1111/1467-9280.00102

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Damasio, H., Tranel, D., Grabowski, T., Adolphs, R., & Damasio, A. (2004). Neural systems behind word and concept retrieval. *Cognition*, *92*(1–2), 179–229. https://doi.org/10.1016/j.cognition.2002.07.001

D'Argembeau, A. (2013). On the role of the ventromedial prefrontal cortex in self-processing: The valuation hypothesis. *Frontiers in Human Neuroscience*, *7*, 1–13. https://doi.org/10.3389/fnhum.2013.00372

D'Argembeau, A. (2018). Mind-wandering and self-referential thought. In K. C. R. Fox & K. Christoff (Eds.), *The Oxford handbook of spontaneous thought* (pp. 181–191). Oxford University Press.

D'Argembeau, A., Collette, F., Van Der Linden, M., Laureys, S., Del Fiore, G., Degueldre, C., Luxen, A., & Salmon, E. (2005). Self-referential reflective activity and its relationship with rest: A PET study. *NeuroImage*, *25*(2), 616–624. https://doi.org/10.1016/j.neuroimage.2004.11.048

D'Argembeau, A., Feyers, D., Majerus, S., Collette, F., Van der Linden, M., Maquet, P., & Salmon, E. (2008). Self-reflection across time: Cortical midline structures differentiate between present and past selves. *Social Cognitive and Affective Neuroscience*, *3*(3), 244–252. https://doi.org/10.1093/scan/nsn020

D'Argembeau, A., Jedidi, H., Balteau, E., Bahri, M., Phillips, C., & Salmon, E. (2012). Valuing one's self: Medial prefrontal involvement in epistemic and emotive investments in self-views. *Cerebral Cortex*, *22*(3), 659–667. https://doi.org/10.1093/cercor/bhr144

D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Balteau, E., Luxen, A., Maquet, P., & Salmon, E. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *Journal of Cognitive Neuroscience*, *19*(6), 935–944. https://doi.org/10.1162/jocn.2007.19.6.935

D'Argembeau, A., Stawarczyk, D., Majerus, S., Collette, F., Van der Linden, M., Feyers, D., & Salmon, E. (2010). The neural basis of personal goal processing when envisioning future events. *Journal of cognitive neuroscience*, *22*(8), 1701-1713. https://doi.org/10.1162/jocn.2009.21314

Davachi, L., Mitchell, J. P., & Wagner, A. D. (2003). Multiple routes to memory: distinct medial temporal lobe processes build item and source memories. *Proceedings of the National Academy of Sciences*, *100*(4), 2157-2162. https://doi.org/10.1073/pnas.0337195100

Davey, C. G., Pujol, J., & Harrison, B. J. (2016). Mapping the self in the brain's default mode network. *NeuroImage*, *132*, 390–397. https://doi.org/10.1016/j.neuroimage.2016.02.022

De Greck, M., Rotte, M., Paus, R., Moritz, D., Thiemann, R., Proesch, U., Bruer, U., Moerth, S., Tempelmann, C., Bogerts, B., & Northoff, G. (2008). Is our self based on reward? Self-relatedness recruits neural activity in the reward system. *NeuroImage*, *39*(4), 2066–2075. https://doi.org/10.1016/j.neuroimage.2007.11.006

Decety, J., & Sommerville, J. A. (2003). Shared representations between self and other: A social cognitive neuroscience view. *Trends in Cognitive Sciences*, *7*(12), 527–533. https://doi.org/10.1016/j.tics.2003.10.004

del Prado, A. M., Timothy Church, A., Katigbak, M. S., Miramontes, L. G., Whitty, M. T., Curtis, G. J., Vargas-Flores, J. de J., Ibáñez-Reyes, J., Ortiz, F. A., & Reyes, J. A. S. (2007). Culture, method, and the content of self-concepts: Testing trait, individual-self-primacy, and cultural psychology perspectives. *Journal of Research in Personality*, *41*(6), 1119–1160. https://doi.org/10.1016/j.jrp.2007.02.002

Delgado, M. R., Beer, J. S., Fellows, L. K., Huettel, S. A., Platt, M. L., Quirk, G. J., & Schiller, D. (2016). Viewpoints: Dialogues on the functional role of the ventromedial prefrontal cortex. *Nature Neuroscience*, *19*(12), 1545–1552. https://doi.org/10.1038/nn.4438

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in

medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *24*(8), 1742–1752. https://doi.org/10.1162/jocn_a_00233

Deutsch, F. M., Ruble, D. N., Fleming, A., Brooks-Gunn, J., & Stangor, C. (1988). Information-Seeking and Maternal Self-Definition During the Transition to Motherhood. *Journal of Personality and Social Psychology*, *55*(3), 420–431. https://doi.org/10.1037/0022-3514.55.3.420

Diehl, M., Jacobs, L. M., & Hastings, C. T. (2006). Temporal stability and authenticity of self-representations in adulthood. *Journal of Adult Development*, *13*(1), 10–22. https://doi.org/10.1007/s10804-006-9001-4

Dimsdale-Zucker, H. R., & Ranganath, C. (2018). Representational Similarity Analyses: A Practical Guide for Functional MRI Applications. *Handbook of Behavioral Neuroscience*, *28*, 509–525. https://doi.org/10.1016/B978-0-12-812028-6.00027-6

Dubois, J., & Adolphs, R. (2016). Building a Science of Individual Differences from fMRI. In *Trends in Cognitive Sciences* (Vol. 20, Issue 6, pp. 425–443). Elsevier Ltd. https://doi.org/10.1016/j.tics.2016.03.014

Dufner, M., Gebauer, J. E., Sedikides, C., & Denissen, J. J. A. (2019). Self-enhancement and psychological adjustment: A meta-analytic review. *Personality and Social Psychology Review*, *23*(1), 48–72. https://doi.org/10.1177/1088868318756467

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*(5643), 290–292. https://doi.org/10.1126/science.1089134

Elder, J., Cheung, B., Davis, T., & Hughes, B. (2022). Mapping the Self: A Network Approach for Understanding Psychological and Neural Representations of Self-Concept Structure. *Journal of Personality and Social Psychology*, *124*(2), 237–263. https://doi.org/10.1037/pspa0000315

Enzi, B., de Greck, M., Prösch, U., Tempelmann, C., & Northoff, G. (2009). Is our self nothing but reward? Neuronal overlap and distinction between reward and personal relevance and its

relation to human personality. *PLoS ONE*, *4*(12), Article e8429. https://doi.org/10.1371/journal.pone.0008429

Esterman, M., Tamber-Rosenau, B. J., Chiu, Y. C., & Yantis, S. (2010). Avoiding non-independence in fMRI data analysis: Leave one subject out. *NeuroImage*, *50*(2), 572–576. https://doi.org/10.1016/j.neuroimage.2009.10.092

Falk, E. B., O'Donnell, M. B., Cascio, C. N., Tinney, F., Kang, Y., Lieberman, M. D., Taylor, S. E., An, L., Resnicow, K., & Strecher, V. J. (2015). Self-affirmation alters the brain's response to health messages and subsequent behavior change. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(7), 1977–1982. https://doi.org/10.1073/pnas.1500247112

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion*, *15*(2), 115–141. https://doi.org/10.1080/02699930125908

Fazio, R. H., & Olson, M. A. (2003). Implicit Measures in Social Cognition Research: Their Meaning and Use. *Annual Review of Psychology*, *54*, 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145225

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Feng, C., Yan, X., Huang, W., Han, S., & Ma, Y. (2018). Neural representations of the multidimensional self in the cortical midline structures. *NeuroImage*, *183*, 291–299. https://doi.org/10.1016/j.neuroimage.2018.08.018

Ferstl, E. C., Neumann, J., Bogler, C., & Von Cramon, D. Y. (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping*, *29*(5), 581–593. https://doi.org/10.1002/hbm.20422

Finn, E. S., Glerean, E., Khojandi, A. Y., Nielson, D., Molfese, P. J., Handwerker, D. A., & Bandettini, P. A. (2020). Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging. *NeuroImage*, *215*. https://doi.org/10.1016/j.neuroimage.2020.116828

Fossati, P., Hevenor, S. J., Graham, S. J., Grady, C., Keightley, M. L., Craik, F., & Mayberg, H. (2003). In search of the emotional self: An fMRI study using positive and negative emotional words. *American Journal of Psychiatry*, *160*(11), 1938–1945. https://doi.org/10.1176/appi.ajp.160.11.1938

Frewen, P., Schroeter, M. L., Riva, G., Cipresso, P., Fairfield, B., Padulo, C., Kemp, A. H., Palaniyappan, L., Owolabi, M., Kusi-Mensah, K., Polyakova, M., Fehertoi, N., D'Andrea, W., Lowe, L., & Northoff, G. (2020). Neuroimaging the consciousness of self: Review, and conceptual-methodological framework. *Neuroscience and Biobehavioral Reviews*, *112*, 164–212. https://doi.org/10.1016/j.neubiorev.2020.01.023

Gainotti, G. (2000). What the locus of brain lesion tells us about the nature of the cognitive defect underlying category-specific disorders: A review. *Cortex*, *36*(4), 539–559. https://doi.org/10.1016/S0010-9452(08)70537-9

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind.' *Trends in Cognitive Sciences*, *7*(2), 77–83. https://doi.org/10.1016/S1364-6613(02)00025-6

Gawronski, B., & Bodenhausen, G. V. (2014). Implicit and explicit evaluation: A brief review of the associative-propositional evaluation model. *Social and Personality Psychology Compass*, *8*(8), 448–462. https://doi.org/10.1111/spc3.12124

Gillihan, S. J., & Farah, M. J. (2005). Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychological Bulletin*, *131*(1), 76–97. https://doi.org/10.1037/0033-2909.131.1.76

Glaser, J., & Banaji, M. R. (1999). When fair is foul and foul is fair: Reverse priming in automatic evaluation. *Journal of Personality and Social Psychology*, *77*(4), 669–687. https://doi.org/10.1037/0022-3514.77.4.669

Goldberg, I. I., Harel, M., & Malach, R. (2006). When the Brain Loses Its Self: Prefrontal Inactivation during Sensorimotor Processing. *Neuron*, *50*(2), 329–339. https://doi.org/10.1016/j.neuron.2006.03.015

Gorgolewski, K. J., & Poldrack, R. A. (2016). A Practical Guide for Improving Transparency and Reproducibility in Neuroimaging Research. *PLoS Biology*, *14*(7). https://doi.org/10.1371/journal.pbio.1002506

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4-27. https://doi.org/10.1037/0033295X.102.1.4

Greenwald, A. G., & Banaji, M. R. (1989). The self as a memory system: Powerful, but ordinary. *Journal of Personality and Social Psychology, 57*(1), 41–54. https://doi.org/10.1037/0022-3514.57.1.41

Greenwald, A. G., & Farnham, S. D. (2000). Using the implicit association test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology*, *79*(6), 1022–1038. https://doi.org/10.1037/0022-3514.79.6.1022

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), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., & Pratkanis, A. R. (1984). The self. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 129–178). Lawrence Erlbaum Associates.

Greenwald, A. G., Rudman, L. A., Nosek, B. A., Banaji, M. R., Farnham, S. D., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*(1), 3–25. https://doi.org/10.1037/0033-295X.109.1.3

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(1), 253–258. https://doi.org/10.1073/pnas.0135058100

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006

Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. *Proceedings of the National Academy of Sciences*, *98*(7), 4259-4264. https://doi.org/10.1073/pnas.071043098

Guthrie, T. D., Benadjaoud, Y. Y., & Chavez, R. S. (2022). Social Relationship Strength Modulates the Similarity of Brain-to-Brain Representations of Group Members. *Cerebral Cortex*, *32*(11), 2469–2477. https://doi.org/10.1093/cercor/bhab355

Haslam, C., Jetten, J., Haslam, S. A., Pugliese, C., & Tonks, J. (2011). "I remember therefore I am, and I am therefore I remember": Exploring the contributions of episodic and semantic self-knowledge to strength of identity. *British Journal of Psychology*, *102*(2), 184–203. https://doi.org/10.1348/000712610X508091

Hassabis, D., Kumaran, D., & Maguire, E. A. (2007). Using imagination to understand the neural basis of episodic memory. *Journal of Neuroscience*, *27*(52), 14365–14374. https://doi.org/10.1523/JNEUROSCI.4549-07.2007

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage*, *62*(2), 852–855. https://doi.org/10.1016/j.neuroimage.2012.03.016

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534. https://doi.org/10.1038/nrn1931

Heatherton, T. F., Wyland, C. L., Macrae, C. N., Demos, K. E., Denny, B. T., & Kelley, W. M. (2006). Medial prefrontal activity differentiates self from close others. *Social Cognitive and Affective Neuroscience*, *1*(1), 18–25. https://doi.org/10.1093/scan/nsl001

Heine, S. J., & Hamamura, T. (2007). In search of east Asian self-enhancement. *Personality and Social Psychology Review*, *11*(1), 4–27. https://doi.org/10.1177/1088868306294587

Heleven, E., & Van Overwalle, F. (2019). The neural representation of the self in relation to close others using fMRI repetition suppression. *Social Neuroscience*, *14*(6), 717–728. https://doi.org/10.1080/17470919.2019.1581657

Hepper, E. G., & Sedikides, C. (2012). Self-enhancing feedback. In R. M. Sutton, M. J. Hornsey, & K. M. Douglas (Eds.), *Feedback: The communication of praise, criticism, and advice* (pp. 43–56). Peter Lang.

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review, 94*(3), 319–340. https://doi.org/10.1037/0033-295X.94.3.319

Hofstadter, D. (2007). *I am a strange loop.* Basic Books.

Huang, A. X., Hughes, T. L., Sutton, L. R., Lawrence, M., Chen, X., Ji, Z., & Zeleke, W. (2017). Understanding the self in individuals with Autism Spectrum Disorders (ASD): A review of literature. *Frontiers in Psychology*, *8*(AUG), 1–8. https://doi.org/10.3389/fpsyg.2017.01422

Huang, C., Sedikides, C., Angus, D. J., Davis, W. E., Butterworth, J. W., Jeffers, A., Schlegel, R. J., & Kelley, N. J. (2025). Demystifying authenticity: Behavioral and neurophysiological signatures of self-positivity for authentic and presented selves. *NeuroImage, 307*, Article 121046. https://doi.org/10.1016/j.neuroimage/2025.121046

Hughes, C., Setton, R., Mwilambwe-Tshilobo, L., Baracchini, G., Turner, G. R., & Spreng, R. N. (2024). Precision mapping of the default network reveals common and distinct (inter) activity for

autobiographical memory and theory of mind. *Journal of Neurophysiology*, *132*(2), 375-388. https://doi.org/10.1152/jn.00427.2023

Iravani, B., Kaboodvand, N., Stieger, J. R., Liang, E. Y., Lusk, Z., Fransson, P., Deutsch, G. K., Gotlib, I. H., & Parvizi, J. (2024). Intracranial recordings of the human orbitofrontal cortical activity during self-referential episodic and valenced self-judgments. *Journal of Neuroscience*, *44*(11), Article e1634232024. https://doi.org/10.1523/JNEUROSCI.1634-23.2024

Iyer, S., Collier, E., Broom, T. W., Finn, E. S., & Meyer, M. L. (2024). Individuals who see the good in the bad engage distinctive default network coordination during post-encoding rest. *Proceedings of the National Academy of Sciences*, *121*(1), Article e2306295121. https://doi.org/10.1073/pnas.2306295121

Izuma, K., Kennedy, K., Fitzjohn, A., Sedikides, C., & Shibata, K. (2018). Neural activity in the reward-related brain regions predicts implicit self-esteem: A novel validity test of psychological measures using neuroimaging. *Journal of Personality and Social Psychology*, *114*(3), 343-357. https://doi.org/10.1037/pspa0000114

James, W. (1890). *The principles of psychology: Volume I*. Henry Holt and Company.

Jenkins, A. C., Macrae, C. N., & Mitchell, J. P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(11), 4507–4512. https://doi.org/10.1073/pnas.0708785105

Jenkins, A. C., & Mitchell, J. P. (2011). Medial prefrontal cortex subserves diverse forms of self-reflection. *Social Neuroscience*, *6*(3), 211–218. https://doi.org/10.1080/17470919.2010.507948

Jimenez, C. A., & Meyer, M. L. (2024). The dorsomedial prefrontal cortex prioritizes social learning during rest. *Proceedings of the National Academy of Sciences*, *121*(12), Article e2309232121. https://doi.org/10.1073/pnas.2309232121

Jolly, E., & Chang, L. J. (2021). Special Issue: Computational Methods in Social Neuroscience Multivariate spatial feature selection in fMRI. *Social Cognitive and Affective Neuroscience, 16*(8), 795–806. https://doi.org/10.1093/scan/nsab010

Kanske, P., Sharifi, M., Smallwood, J., Dziobek, I., & Singer, T. (2017). Where the narcissistic mind wanders: Increased Self-Related thoughts are more positive and future oriented. *Journal of Personality Disorders*, *31*(4), 553–566. https://doi.org/10.1521/pedi_2016_30_263

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*, Article 151. https://doi.org/10.3389/fnhum.2015.00151

Kawakami, K., & Dovidio, J. F. (2001). The reliability of implicit stereotyping. *Personality and Social Psychology Bulletin*, *27*(2), 212-225. https://doi.org/10.1177/0146167201272007

Kelley, A. W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience*, *14*(5), 785–794. https://doi.org/10.1162/08989290260138672

Kihlstrom, J. F., & Cantor, N. (1984). Mental representations of the self. *Advances in Experimental Social Psychology*, *17*, 1–47. https://doi.org/10.1016/S0065-2601(08)60117-3

Kihlstrom, J. F., Beer, J. S., & Klein, S. B. (2003). Self and identity as memory. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 68–90). Guilford Press.

Kihlstrom, J. F., & Klein, S. B. (1994). The self as a knowledge structure. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 153–208). Lawrence Erlbaum Associates.

Kim, H. (2012). A dual-subsystem model of the brain's default network: Self-referential processing, memory retrieval processes, and autobiographical memory retrieval. *NeuroImage*, *61*(4), 966–977. https://doi.org/10.1016/j.neuroimage.2012.03.025

Kim, K., & Johnson, M. K. (2012). Extended self: Medial prefrontal activity during transient association of self and objects. *Social Cognitive and Affective Neuroscience*, *7*(2), 199–207. https://doi.org/10.1093/scan/nsq096

Klauer, K. C. (1997). Affective Priming. *European Review of Social Psychology*, *8*(1), 67–103. https://doi.org/10.1080/14792779643000083

Klein, S. B., Babey, S. H., & Sherman, J. W. (1997). The functional independence of trait and behavioral self-knowledge: Methodological considerations and new empirical findings. *Social Cognition*, *15*(3), 183–203. https://doi.org/10.1521/soco.1997.15.3.183

Klein, S. B., Cosmides, L., & Costabile, K. A. (2003). Preserved knowledge of self in a case of Alzheimer's dementia. *Social Cognition*, *21*(2), 157-165. https://doi.org/10.1521/soco.21.2.157.21317

Klein, S. B., & Kihlstrom, J. F. (1986). Elaboration, Organization, and the Self-Reference Effect in Memory. *Journal of Experimental Psychology: General*, *115*(1), 26–38. https://doi.org/10.1037/0096-3445.115.1.26

Klein, S. B., & Lax, M. L. (2010). The unanticipated resilience of trait self-knowledge in the face of neural damage. *Memory*, *18*(8), 918–948. https://doi.org/10.1080/09658211.2010.524651

Klein, S. B., & Loftus, J. (1988). The nature of self-referent encoding: The contributions of elaborative and organizational processes. *Journal of Personality and Social Psychology*, *55*(1), 5–11. https://doi.org/10.1037/0022-3514.55.1.5

Klein, S. B., & Loftus, J. (1990). Rethinking the role of organization in person memory: An independent trace storage model. *Journal of Personality and Social Psychology*, *59*(3), 400–410. https://doi.org/10.1037/0022-3514.59.3.400

Klein, S. B., & Loftus, J. (1993). The mental representation of trait and autobiographical knowledge about the self. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *The mental representation of trait and autobiographical knowledge about the self* (pp. 1–49). Lawrence Erlbaum Associates.

Klein, S. B., Loftus, J., & Burton, H. A. (1989). Two Self-Reference Effects: The Importance of distinguishing between self-descriptiveness judgments and autobiographical retrieval in self-referent encoding. *Journal of Personality and Social Psychology*, *56*(6), 853–865. https://doi.org/10.1037/0022-3514.56.6.853

Klein, S. B., Loftus, J., & Kihlstrom, J. F. (1996). Self-knowledge of an amnesic patient: Toward a neuropsychology of personality and social psychology. *Journal of Experimental Psychology: General, 125*(3), 250–260. https://doi.org/10.1037/0096-3445.125.3.250

Klein, S. B., Loftus, J., & Kihlstrom, J. F. (2002). Memory and temporal experience: The effects of episodic memory loss on an amnesic patient's ability to remember the past and imagine the future. *Social Cognition*, *20*(5), 353–379. https://doi.org/10.1521/soco.20.5.353.21125

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140–151. https://doi.org/10.1016/j.neuron.2008.11.027

Koban, L., Gianaros, P. J., Kober, H., & Wager, T. D. (2021). The self in context: brain systems linking mental and physical health. *Nature Reviews Neuroscience*, *22*(5), 309–322. https://doi.org/10.1038/s41583-021-00446-8

Konishi, M., McLaren, D. G., Engen, H., & Smallwood, J. (2015). Shaped by the past: The default mode network supports cognition that is independent of immediate perceptual input. *PLoS ONE*, *10*(6), 1–18. https://doi.org/10.1371/journal.pone.0132209

Koole, S., & Coenen, L. (2007). Implicit self and affect regulation: Effects of action orientation and subliminal self priming in an affective priming task. *Self and Identity*, *6*(3–4), 118–136. https://doi.org/10.1080/15298860601118835

Koole, S. L., Dijksterhuis, A., & Van Knippenberg, A. D. (2001). What's in a name: implicit self-esteem and the automatic self. *Journal of Personality and Social Psychology*, *80*(4), 669-685. https://doi.org/10.1037/0022-3514.80.4.669

Koski, J. E., McHaney, J. R., Rigney, A. E., & Beer, J. S. (2020). Reconsidering longstanding assumptions about the role of medial prefrontal cortex (MPFC) in social evaluation. *NeuroImage*, *214*,*Article 116752*. https://doi.org/10.1016/j.neuroimage.2020.116752

Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., & Saxe, R. (2017). Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage*, *161*, 9–18. https://doi.org/10.1016/j.neuroimage.2017.08.026

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, Article 4. https://doi.org/10.3389/neuro.06.004.2008

Krienen, F. M., Tu, P. C., & Buckner, R. L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience*, *30*(41), 13906–13915. https://doi.org/10.1523/JNEUROSCI.2180-10.2010

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

Kuhn, M. H., & McPartland, T. S. (1954). Twenty statements test. *American Sociological Review, 19*(1), 68-76. https://doi.org/10.1037/t05100-000

Kuiper, N. A., & Rogers, T. B. (1979). Encoding of personal information: Self–other differences. *Journal of Personality and Social Psychology*, *37*(4), 499–514. https://doi.org/10.1037/0022-3514.37.4.499

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the Open Affective Standardized Image Set (OASIS). *Behavior Research Methods*, *49*(2), 457–470. https://doi.org/10.3758/s13428-016-0715-3

Legrand, D., & Ruby, P. (2009). What is self-specific? Theoretical investigation and critical review of neuroimaging results. *Psychological Review*, *116*(1), 252–282. https://doi.org/10.1037/a0014172

Levorsen, M., Aoki, R., Matsumoto, K., Sedikides, C., & Izuma, K. (2023). The self-concept Is represented in the medial prefrontal cortex in terms of self-importance. *Journal of Neuroscience*, *43*(20), 3675–3686. https://doi.org/10.1523/JNEUROSCI.2178-22.2023

Levorsen, M., Ito, A., Suzuki, S., & Izuma, K. (2021). Testing the reinforcement learning hypothesis of social conformity. *Human Brain Mapping*, *42*(5), 1328–1342. https://doi.org/10.1002/hbm.25296

Lin, W. J., Horner, A. J., & Burgess, N. (2016). Ventromedial prefrontal cortex, adding value to autobiographical memories. *Scientific Reports*, *6*, Article 28630. https://doi.org/10.1038/srep28630

Linville, P. W. (1985). Self-complexity and affective extremity: Don't put all of your eggs in one cognitive basket. *Social cognition*, *3*(1), 94-120. https://doi.org/10.1521/soco.1985.3.1.94

Linville, P. W. (1987). Self-complexity as a cognitive buffer against stress-related illness and depression. *Journal of Personality and Social Psychology, 52*(4), 663–676. https://doi.org/10.1037/0022-3514.52.4.663

Locksley, A., & Lenauer, M. (1981). Considerations for a theory of self-inference processes. In N. Cantor & J. K. Kihlstrom (Eds.), *Personality, cognition, and social interaction* (pp. 263–277). Routledge.

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Sadek, S. A., Pasco, G., Wheelwright, S. J., Suckling, J., & Baron-Cohen, S. (2010). Atypical neural self-representation in autism. *Brain*, *133*(2), 611–624. https://doi.org/10.1093/brain/awp306

Lou, H. C., Luber, B., Crupain, M., Keenan, J. P., Nowak, M., Kjaer, T. W., Sackeim, H. A., & Lisanby, S. H. (2004). Parietal cortex and representation of the mental Self. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(17), 6827–6832. https://doi.org/10.1073/pnas.0400049101

Lu, H., Zou, Q., Gu, H., Raichle, M. E., Stein, E. A., & Yang, Y. (2012). Rat brains also have a default mode network. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(10), 3979–3984. https://doi.org/10.1073/pnas.1200506109

Macrae, C. N., Moran, J. M., Heatherton, T. F., Banfield, J. F., & Kelley, W. M. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex*, *14*(6), 647–654. https://doi.org/10.1093/cercor/bhh025

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

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 35*(2), 63–78. https://doi.org/10.1037/0022-3514.35.2.63

Markus, H., Crane, M., Bernstein, S., & Siladi, M. (1982). Self-schemas and gender. *Journal of Personality and Social Psychology*, *42*(1), 38-50. https://doi.org/10.1037/0022-3514.42.1.38

Markus, H. (1983). Self-knowledge: An expanded view. *Journal of Personality*, *51*(3), 543-565. https://doi.org/10.1111/j.1467-6494.1983.tb00344.x

Markus, H., & Sentis, K. (1982). The self in social information processing. In J. Suls (Ed.), *Social psychological perspectives on the self* (pp. 41-70). Lawrence Erlbaum.

Markus, H., & Wurf, E. (1987). The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology, 38*, 299-337.https://doi.org/10.1146/annurev.ps.38.020187.001503

Marquine, M. J., Grilli, M. D., Rapcsak, S. Z., Kaszniak, A. W., Ryan, L., Walther, K., & Glisky, E. L. (2016). Impaired personal trait knowledge, but spared other-person trait knowledge, in an individual with bilateral damage to the medial prefrontal cortex. *Neuropsychologia*, *89*, 245–253. https://doi.org/10.1016/j.neuropsychologia.2016.06.021

Martial, C., Stawarczyk, D., & D'Argembeau, A. (2018). Neural correlates of context-independent and context-dependent self-knowledge. *Brain and Cognition*, *125*, 23–31. https://doi.org/10.1016/j.bandc.2018.05.004

Martinelli, P., Sperduti, M., & Piolino, P. (2013). Neural substrates of the self-memory system: New insights from a meta-analysis. *Human Brain Mapping*, *34*(7), 1515–1529. https://doi.org/10.1002/hbm.22008

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*(4), 115–133. https://doi.org/10.1007/BF02478259

McConnell, A. R., Renaud, J. M., Dean, K. K., Green, S. P., Lamoreaux, M. J., Hall, C. E., & Rydell, R. J. (2005). Whose self is it anyway? Self-aspect control moderates the relation between self-complexity and well-being. *Journal of Experimental Social Psychology*, *41*(1), 1–18. https://doi.org/10.1016/j.jesp.2004.02.004

McConnell, A. R., & Strain, L. M. (2011). Content and structure of the self-concept. In C. Sedikides & S. Spencer (Eds.), *The self in social psychology* (pp. 51–74). Psychology Press. https://doi.org/10.4324/9780203818572

McGuire, W. J., & McGuire, C. V. (1982). Significant others in self-space: Sex differences and developmental trends in the social self. In J. Suls & A. G. Greenwald (Eds.), *Psychological perspectives on the self* (Vol. 1, pp. 71–96). Lawrence Erlbaum Associates.

Menon, V. (2023). 20 years of the default mode network: A review and synthesis. *Neuron*, *111*(16), 2469–2487. https://doi.org/10.1016/j.neuron.2023.04.023

Metzinger, T. (2009). *The ego tunnel: The science of the mind and the myth of the self*. Basic Books.

Meyer, M. L., Davachi, L., Ochsner, K. N., & Lieberman, M. D. (2019). Evidence that default network connectivity during rest consolidates social information. *Cerebral Cortex*, *29*(5), 1910–1920. https://doi.org/10.1093/cercor/bhy071

Meyer, M. L., & Lieberman, M. D. (2018). Why people are always thinking about themselves: medial prefrontal cortex activity during rest primes self-referential processing. *Journal of Cognitive Neuroscience*, *30*(5), 714-721. https://doi.org/10.1162/jocn_a_01232

Midgley, M. (2014). *Are you an illusion?* Routledge.

Misaki, M., Kerr, K. L., Ratliff, E. L., Cosgrove, K. T., Simmons, W. K., Morris, A. S., & Bodurka, J. (2021). Beyond synchrony: The capacity of fMRI hyperscanning for the study of human social interaction. *Social Cognitive and Affective Neuroscience*, *16*(1–2), 84–92. https://doi.org/10.1093/scan/nsaa143

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *17*(8), 1306–1315. https://doi.org/10.1162/0898929055002418

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, *50*(4), 655–663. https://doi.org/10.1016/j.neuron.2006.03.040

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews*, *65*, 276–291. https://doi.org/10.1016/j.neubiorev.2016.03.020

Moran, J. M., Heatherton, T. F., & Kelley, W. M. (2009). Modulation of cortical midline structures by implicit and explicit self-relevance evaluation. *Social Neuroscience*, *4*(3), 197–211. https://doi.org/10.1080/17470910802250519

Moran, J. M., Lee, S. M., & Gabrieli, J. D. E. (2011). Dissociable neural systems supporting knowledge about human character and appearance in ourselves and others. *Journal of Cognitive Neuroscience*, *23*(9), 2222–2230. https://doi.org/10.1162/jocn.2010.21580

Moran, J.M., Macrae, C.N., Heatherton, T.F., Wyland, C.L., Kelley, W.M. (2006). Neuroanatomical evidence for distinct cognitive and affective components of self. *Journal of Cognitive Neuroscience, 18*, 1586-1594. https://doi.org/10.1162/jocn.2006.18.9.1586

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17-23. https://doi.org/10.2307/2332142

Mumford, J. A., Poline, J. B., & Poldrack, R. A. (2015). Orthogonalization of regressors in fMRI models. *PLoS ONE*, *10*(4), Article e0126255. https://doi.org/10.1371/journal.pone.0126255

Murray, R. J., Schaer, M., & Debbané, M. (2012). Degrees of separation: A quantitative neuroimaging meta-analysis investigating self-specificity and shared neural activation between self- and other-reflection. *Neuroscience and Biobehavioral Reviews*, *36*(3), 1043–1059. https://doi.org/10.1016/j.neubiorev.2011.12.013

Nakao, T., Ohira, H., & Northoff, G. (2012). Distinction between externally vs. Internally guided decision-making: Operational differences, meta-analytical comparisons and their theoretical implications. *Frontiers in Neuroscience*, *6*, Article 31. https://doi.org/10.3389/fnins.2012.00031

Neininger, B., & Pulvermüller, F. (2003). Word-category specific deficits after lesions in the right hemisphere. *Neuropsychologia*, *41*(1), 53–70.https://doi.org/10.1016/S0028-3932(02)00126-4

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Prentice-Hall.

Nichols, T., & Holmes, A. (2003). Nonparametric permutation tests for functional neuroimaging. In R. S. J. Frackowiak, J. T. Ashburner, & K. J. Friston (Eds.), *Human brain function* (2nd ed., pp. 887–910). Academic Press.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4), Article e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

https://doi.org/10.1037/0033-295X.84.3.231

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

Northoff, G. (2016). Is the self a higher-order or fundamental function of the brain? The "basis model of self-specificity" and its encoding by the brain's spontaneous activity. *Cognitive Neuroscience*, *7*(1–4), 203–222. https://doi.org/10.1080/17588928.2015.1111868

Northoff, G., & Hayes, D. J. (2011). Is our self nothing but reward? *Biological Psychiatry*, *69*(11), 1019–1025. https://doi.org/10.1016/j.biopsych.2010.12.014

Northoff, G., Schneider, F., Rotte, M., Matthiae, C., Tempelmann, C., Wiebking, C., Bermpohl, F., Heinzel, A., Danos, P., Heinze, H. J., Bogerts, B., Walter, M., & Panksepp, J. (2009). Differential parametric modulation of self-relatedness and emotions in different brain regions. *Human Brain Mapping*, *30*(2), 369–382. https://doi.org/10.1002/hbm.20510

Ochsner, K. N., Beer, J. S., Robertson, E. R., Cooper, J. C., Gabrieli, J. D. E., Kihsltrom, J. F., & D'Esposito, M. (2005). The neural correlates of direct and reflected self-knowledge. *NeuroImage,* 28(4), 797-814. https://doi.org/10.1016/j.neuroimage.2005.06.069

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT Press.

Ostrom, T. M., Lingle, J. H., Pryor, J. B., & Geva, N. (1980). Cognitive organization of person impressions. In R. Hastie, T. M. Ostrom, E. B. Ebbesen, R. S. Wyer, D. L. Hamilton, & D. E. Carlston (Eds.), *Person memory: The cognitive basis of social perception* (pp. 55–88). Erlbaum.

Panayiotou, G., & Vrana, S. R. (2004). The role of self-focus, task difficulty, task self-relevance, and evaluation anxiety in reaction time performance. *Motivation and Emotion*, *28*(2), 171-196. https://doi.org/10.1023/B:MOEM.0000032313.69675.0d

Parelman, J. M., Dore, B. P., Cooper, N., O'Donnell, M. B., Chan, H. Y., & Falk, E. B. (2022). Overlapping Functional Representations of Self-and Other-Related Thought are Separable Through Multivoxel Pattern Classification. *Cerebral Cortex*, *32*(6), 1131–1141. https://doi.org/10.1093/cercor/bhab272

Parvizi, J., & Kastner, S. (2018). Promises and limitations of human intracranial electroencephalography. *Nature Neuroscience*, *21*(4), 474–483. https://doi.org/10.1038/s41593-018-0108-2

Phan, K. L., Taylor, S. F., Welsh, R. C., Ho, S. H., Britton, J. C., & Liberzon, I. (2004). Neural correlates of individual ratings of emotional salience: A trial-related fMRI study. *NeuroImage*, *21*(2), 768–780. https://doi.org/10.1016/j.neuroimage.2003.09.072

Philippi, C. L., Duff, M. C., Denburg, N. L., Tranel, D., & Rudrauf, D. (2012). Medial PFC damage abolishes the self-reference effect. *Journal of Cognitive Neuroscience*, *24*(2), 475-481. https://doi.org/10.1162/jocn_a_00138

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2020). The present and future use of functional near-infrared spectroscopy (Fnirs) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1464*(1), 5–29. https://doi.org/10.1111/nyas.13948

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J. B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*(2), 115–126. https://doi.org/10.1038/nrn.2016.167

Popal, H., Wang, Y., & Olson, I. R. (2019). A Guide to Representational Similarity Analysis for Social Neuroscience. *Social Cognitive and Affective Neuroscience*, *14*(11), 1243–1253. https://doi.org/10.1093/scan/nsz099

Popov, L. M., & Ilesanm, R. A. (2015). Parent-child relationship: Peculiarities and outcome. *Review of European Studies*, *7*(5), 253–263. https://doi.org/10.5539/res.v7n5p253

Qin, P., & Northoff, G. (2011). How is our self related to midline regions and the default-mode network? *NeuroImage*, *57*(3), 1221–1233. https://doi.org/10.1016/j.neuroimage.2011.05.028

Raichle, M. E. (2015). The Brain's Default Mode Network. *Annual Review of Neuroscience*, *38*, 433–447. https://doi.org/10.1146/annurev-neuro-071013-014030

Rameson, L. T., Satpute, A. B., & Lieberman, M. D. (2010). The neural correlates of implicit and explicit self-relevant processing. *NeuroImage*, *50*(2), 701-708. https://doi.org/10.1016/j.neuroimage.2009.12.098

Reed II, A., & Aquino, K. F. (2003). Moral identity and the expanding circle of moral regard toward out-groups. *Journal of Personality and Social Psychology*, *84*(6), 1270-1286. https://doi.org/10.1037/0022-3514.84.6.1270

Ritchie, J. B., Bracci, S., & Op de Beeck, H. (2017). Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *NeuroImage*, *148*, 197–200. https://doi.org/10.1016/j.neuroimage.2016.12.079

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-Reference and the Encoding of Personal Information. In *Journal of Personality and Social Psychology*, *35*(9), 677–688. https://doi.org/10.1037/0022-3514.35.9.677

Rosenberg, M. (1979). Conceiving the Self. Basic Books.

Ruby, P. and Legrand, D. (2007) Neuroimaging the self? In Y. Rossetti, P. Haggard and M. Kawato (Eds.) *Sensorimotor Foundations of Higher Cognition* (pp. 293-318). Oxford University Press.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, *19*(2), 65–72. https://doi.org/10.1016/j.tics.2014.11.007

Schmitz, T. W., & Johnson, S. C. (2007). Relevance to self: A brief review and framework of neural systems underlying appraisal. *Neuroscience and Biobehavioral Reviews*, *31*(4), 585–596. https://doi.org/10.1016/j.neubiorev.2006.12.003

Schmitz, T. W., Kawahara-Baccus, T. N., & Johnson, S. C. (2004). Metacognitive evaluation, self-relevance, and the right prefrontal cortex. *NeuroImage*, *22*(2), 941–947. https://doi.org/10.1016/j.neuroimage.2004.02.018

Schulz, S. M. (2016). Neural correlates of heart-focused interoception: A functional magnetic resonance imaging meta-analysis. *Philosophical Transactions of the Royal Society: Biological Sciences*, *371*(1708). https://doi.org/10.1098/rstb.2016.0018

Sedikides, C. (1993). Assessment, enhancement, and verification determinants of the self-evaluation process. *Journal of Personality and Social Psychology*, *65*(2), 317-338. https://doi.org/10.1037/0022-3514.65.2.317

Sedikides, C. (1995). Central and peripheral self-conceptions are differentially influenced by mood: tests of the differential sensitivity hypothesis. *Journal of Personality and Social psychology*, *69*(4), 759-779. https://doi.org/10.1037/0022-3514.69.4.759

Sedikides, C. (2020). On the doggedness of self-enhancement and self-protection: How constraining are reality constraints? *Self and Identity, 19*(3), 251-271. https://doi.org/10.1080/15298868.2018.1562961

Sedikides, C., Alicke, M. D., & Skowronski, J. J. (2021). On the utility of the self in social perception: An Egocentric Tactician Model. *Advances in Experimental Social Psychology, 63*(1), 247–298. https://doi.org/10.1016/bs.aesp.2020.11.005

Sedikides, C., & Green, J. D. (2000). On the self-protective nature of inconsistency/negativity management: Using the person memory paradigm to examine self-referent memory. *Journal of Personality and Social Psychology, 79*(6), 906–922. https://doi.org/10.1037/0022-3514.79.6.906

Sedikides, C., Green, J. D., Saunders, J., Skowronski, J. J., & Zengel, B. (2016). Mnemic neglect: Selective amnesia of one's faults. *European Review of Social Psychology*, *27*(1), 1–62. https://doi.org/10.1080/10463283.2016.1183913

Sedikides, C., & Gregg, A. P. (2003). Portraits of the self. In M. A. Hogg & J. Cooper (Eds.), *Sage handbook of social psychology* (pp. 110–138). Sage Publications. https://doi.org/10.4135/9781848608221.n5

Sedikides, C., & Gregg, A.P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science, 3*(2), 102-116. https://doi.org/10.1111/j.1745-6916.2008.00068.x

Sedikides, C., & Skowronski, J. J. (1997). The symbolic self in evolutionary context. *Personality and Social Psychology Review*, *1*(1), 80-102. https://doi.org/10.1207/s15327957pspr0101_6

Sedikides, C., & Skowronski, J. J. (2000). On the evolutionary functions of the symbolic self: The emergence of self-evaluation motives. https://doi.org/10.1037/10357-004

Sedikides, C., Skowronski, J. J., & Dunbar, R. I. (2006). When and why did the human self evolve? In Schaller, M., Simpson, J.A., & Kenrick, D.T. (Eds.), *Evolution and social psychology* (pp. 55-80). Psychology Press.

Sedikides, C., & Spencer, S. (2007). *The self: Frontiers in social psychology*. Psychology Press.

Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. *Advances in Experimental Social Psychology*, *29*, 209–269. https://doi.org/10.1016/S0065-2601(08)60018-0

Segal, Z. V., Hood, J. E., Shaw, B. F., & Higgins, E. T. (1988). A structural analysis of the self-schema construct in major depression. *Cognitive Therapy and Research*, *12*(5), 471–485. https://doi.org/10.1007/BF01173414

Seger, C. A., Stone, M., & Keenan, J. P. (2004). Cortical Activations during judgments about the self and an other person. *Neuropsychologia*, *42*(9), 1168–1177. https://doi.org/10.1016/j.neuropsychologia.2004.02.003

Seitz, R. J., Franz, M., & Azari, N. P. (2009). Value judgments and self-control of action: The role of the medial frontal cortex. *Brain Research Reviews*, *60*(2), 368–378. https://doi.org/10.1016/j.brainresrev.2009.02.003

Shulman, G. L., Corbetta, M., Buckner, R. L., Fiez, J. A., Miezin, F. M., Raichle, M. E., & Petersen, S. E. (1997). Common blood flow changes across visual tasks: I. Increases in subcortical structures and cerebellum but not in nonvisual cortex. *Journal of Cognitive Neuroscience*, *9*(5), 624–647. https://doi.org/10.1162/jocn.1997.9.5.624

Skowronski, J.J., Walker, W.R., Henderson, D.X., & Bond, G.D. (2014). The fading affect bias: Its history, its implications, and its future. *Advances in Experimental Social Psychology, 49*, 163–218. https://doi.org/10.1016/B978-0-12-800052-6.00003-2

Smith, J. F., Hur, J., Kaplan, C. M.,& Shackman, A. J. (2018) *The impact of spatial normalization for functional magnetic resonance imaging data analyses revisited. bioRxiv*. https://doi.org/10.1101/272302

Soares, A. P., Macedo, J., Oliveira, H. M., Lages, A., Hernández-Cabrera, J., & Pinheiro, A. P. (2019). Self-reference is a fast-acting automatic mechanism on emotional word processing: evidence

from a masked priming affective categorisation task. *Journal of Cognitive Psychology*, *31*(3), 317–325. https://doi.org/10.1080/20445911.2019.1599003

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience*, *21*(3), 489–510. https://doi.org/10.1162/jocn.2008.21029

Stafford, J. M., Jarrett, B. R., Miranda-Dominguez, O., Mills, B. D., Cain, N., Mihalas, S., Lahvis, G. P., Lattal, K. M., Mitchell, S. H., David, S. V., Fryer, J. D., Nigg, J. T., & Fair, D. A. (2014). Large-scale topology and the default mode network in the mouse connectome. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(52), 18745–18750. https://doi.org/10.1073/pnas.1404346111

Stawarczyk, D., Majerus, S., Maj, M., Van der Linden, M., & D'Argembeau, A. (2011). Mind-wandering: Phenomenology and function as assessed with a novel experience sampling method. *Acta Psychologica*, *136*(3), 370–381. https://doi.org/10.1016/j.actpsy.2011.01.002

Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Cortical midline involvement in autobiographical memory. *NeuroImage*, *44*(3), 1188–1200. https://doi.org/10.1016/j.neuroimage.2008.09.033

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: a meta-analysis. *Psychological Bulletin*, *121*(3), 371. https://doi.org/10.1037/0033-2909.121.3.371

Szucs, D., & Ioannidis, J. P. (2020). Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *NeuroImage*, *221*. https://doi.org/10.1016/j.neuroimage.2020.117164

Tamir, D. I., & Mitchell, J. P. (2011). The default network distinguishes construals of proximal versus distal events. *Journal of Cognitive Neuroscience*, *23*(10), 2945–2955. https://doi.org/10.1162/jocn_a_00009

Tan, K. M., Daitch, A. L., Pinheiro-Chagas, P., Fox, K. C. R., Parvizi, J., & Lieberman, M. D. (2022). Electrocorticographic evidence of a common neurocognitive sequence for mentalizing about the self and others. *Nature Communications*, *13*(1), 1–17. https://doi.org/10.1038/s41467-022-29510-2

Thornton, M. A., & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex*, *28*(10), 3505–3520. https://doi.org/10.1093/cercor/bhx216

Tsoi, L., Burns, S. M., Falk, E. B., & Tamir, D. I. (2022). The promises and pitfalls of functional magnetic resonance imaging hyperscanning for social interaction research. *Social and Personality Psychology Compass*, *16*(10), 1–20. https://doi.org/10.1111/spc3.12707

Tulving, E. (1983). Ecphoric processes in episodic memory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *302*(1110), 361-371. https://doi.org/10.1098/rstb.1983.0060

Tulving, E. (1993). What is episodic memory? *Current Directions in Psychological Science*, *2*(3), 67-70. https://doi.org/10.1111/1467-8721.ep10770899

Tulving, E., Schacter, D. L., Mclachlan, D. R., & Moscovitch, M. (1988). Priming of semantic autobiographical knowledge: A case study of retrograde amnesia. *Brain and Cognition*, *8*(1), 3-20. https://doi.org/10.1016/0278-2626(88)90035-8

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*(1). https://doi.org/10.1038/s42003-018-0073-z

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352. https://doi.org/10.1037/0033-295X.84.4.327

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist*, *30*(2), 128. https://doi.org/10.1037/h0076759

Vaidya, A. R., Pujara, M. S., Petrides, M., Murray, E. A., & Fellows, L. K. (2019). Lesion studies in contemporary neuroscience. *Trends in Cognitive Sciences*, *23*(8), 653–671. https://doi.org/10.1016/j.tics.2019.05.009

Van Buuren, M., Sijtsma, H., Lute, N., van Rijn, R., Hollarek, M., Walsh, R. J., Lee, N. C., & Krabbendam, L. (2022). Development of the neural correlates of self- and other-referential processing across adolescence. *NeuroImage*, *252*, Article 119032. https://doi.org/10.1016/j.neuroimage.2022.119032

Van der Meer, L., Costafreda, S., Aleman, A., & David, A. S. (2010). Self-reflection and the brain: A theoretical review and meta-analysis of neuroimaging studies with implications for schizophrenia. *Neuroscience and Biobehavioral Reviews*, *34*(6), 935–946. https://doi.org/10.1016/j.neubiorev.2009.12.004

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858. https://doi.org/10.1002/hbm.20547

Vanderwal, T., Hunyadi, E., Grupe, D. W., Connors, C. M., & Schultz, R. T. (2008). Self, mother and abstract other: An fMRI study of reflective social processing. *NeuroImage*, *41*(4), 1437–1446. https://doi.org/10.1016/j.neuroimage.2008.03.058

Van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. https://doi.org/10.1016/j.jesp.2016.03.004

Vincent, J. L., Patel, G. H., Fox, M. D., Snyder, A. Z., Baker, J. T., Van Essen, D. C., Zempel, J. M., Snyder, L. H., Corbetta, M., & Raichle, M. E. (2007). Intrinsic functional architecture in the anaesthetized monkey brain. *Nature*, *447*(7140), 83–86. https://doi.org/10.1038/nature05758

Wagner, D. D., Chavez, R. S., & Broom, T. W. (2019). Decoding the neural representation of self and person knowledge with multivariate pattern analysis and data-driven approaches. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(1), 1–19. https://doi.org/10.1002/wcs.1482

Wagner, D. D., Haxby, J. V., & Heatherton, T. F. (2012). The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(4), 451–470. https://doi.org/10.1002/wcs.1183

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), 1558–1564. https://doi.org/10.1093/scan/nsx092

Wen, T., Mitchell, D. J., & Duncan, J. (2020). The Functional Convergence and Heterogeneity of Social, Episodic, and Self-Referential Thought in the Default Mode Network. *Cerebral Cortex*, *30*(11), 5915–5929. https://doi.org/10.1093/cercor/bhaa166

Wheatley, T., Thornton, M. A., Stolk, A., & Chang, L. J. (2024). The emerging science of interacting minds. *Perspectives on Psychological Science*, *19*(2), 355–373. https://doi.org/10.1177/17456916231200177

Wicker, B., Ruby, P., Royet, J. P., & Fonlupt, P. (2003). A relation between rest and the self in the brain? *Brain Research Reviews*, *43*(2), 224–230. https://doi.org/10.1016/j.brainresrev.2003.08.003

Winograd, T. (1975). Frame representations and the declarative-procedural controversy. In D. Bobrow & A. Collins (Eds.), *Representation and understanding: Studies in cognitive science* (pp. 185–210). Academic Press. https://doi.org/10.1016/b978-0-12-108550-6.50012-4

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

Yankouskaya, A., Humphreys, G., Stolte, M., Stokes, M., Moradi, Z., & Sui, J. (2017). An anterior-posterior axis within the ventromedial prefrontal cortex separates self and reward. *Social Cognitive and Affective Neuroscience*, *12*(12), 1859–1868. https://doi.org/10.1093/scan/nsx112

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665-670. https://doi.org/10.1038/nmeth.1635

Yaoi, K., Osaka, M., & Osaka, N. (2015). Neural correlates of the self-reference effect: Evidence from evaluation and recognition processes. *Frontiers in Human Neuroscience*, *9*, Article 383. https://doi.org/10.3389/fnhum.2015.00383

Yeshurun, Y., Nguyen, M., & Hasson, U. (2021). The default mode network: where the idiosyncratic self meets the shared social world. *Nature Reviews Neuroscience*, *22*(3), 181–192. https://doi.org/10.1038/s41583-020-00420-w

Zhang, M., Jia, H., Zheng, M., & Liu, T. (2021). Group decision-making behavior in social dilemmas: Inter-brain synchrony and the predictive role of personality traits. *Personality and Individual Differences*, *168*, Article 110315. https://doi.org/10.1016/j.paid.2020.110315

Zhang, L., Zhou, T., Zhang, J., Liu, Z., Fan, J., & Zhu, Y. (2006). In search of the Chinese self: An fMRI study. *Science in China, Series C: Life Sciences*, *49*(1), 89–96. https://doi.org/10.1007/s11427-004-5105-x

Zhu, L., Guo, X., Li, J., Zheng, L., Wang, Q., & Yang, Z. (2012). Hippocampal activity is associated with self-descriptiveness effect in memory, whereas self-reference effect in memory depends on medial prefrontal activity. *Hippocampus*, *22*(7), 1540–1552. https://doi.org/10.1002/hipo.20994

Zhu, Y., Zhang, L., Fan, J., & Han, S. (2007). Neural basis of cultural influence on self-representation. *NeuroImage*, *34*(3), 1310–1316. https://doi.org/10.1016/j.neuroimage.2006.08.047