Special thank you to Jason Mitchell, Ruben van Genugten and Nicholas Harkness

semantic representations of self and others in the theory of mind network

Emilė Radytė | Supervised by Mitchell (Psychology) & Harkness (Anthropology)

Abstract

Humans navigate social environments daily: understanding the minds and thoughts of others, known as mentalizing, is crucial to performing this activity. Recent work in social cognitive neuroscience has suggested that the brain may have a cognitive architecture to process complex information. In this thesis, I study whether a specific architecture that is used to process information about other people's thoughts can be found, and if that architecture remains stable when "mentalizing" about oneself. In order to conceptualize this system of explicit neural representations of cognitive processes, we used a linguistic metaphor of "semantic" relations between "agent" (the person having the thought) and the "patient" (the target of the thought). Our experiments used functional magnetic resonance imaging (fMRI), where participants are prompted to engage with the thoughts of other personally familiar people over a series of adapted scenarios that either include themselves (self-others task) or do not include themselves (other people task). We interpreted collected neuroimaging data using representational similarity analysis (RSA). Preliminary results do not support the semantic metaphor of "agent" and "patient" for neural representations. However, while distinct role-based categories are unlikely, there may be emergent representations in the bound relation between role and thought, suggesting directions for further research. I use the empirical findings of linguistic anthropology to highlight and explain the more, and less, promising parts of the cognitive theories and linguistic metaphors at the core of this research and reflect on the methodology through an interdisciplinary lens that could lead to future improvements in research design.

Background

There are important parallels between social cognition and neural activity of the brain. This thesis uses insights from linguistic anthropology and semiotics to investigate whether mental representations, as understood within the framework of *compositionality* and *language of thought* hypotheses, exist when mentalizing (thinking about the minds of other people), and can be studied within the empirical framing of agent-patient relations.

Social cognition	Neural representation
People thinking about other people e.g. I am thinking about someone named Ife and her thoughts, as opposed to thinking about a chair.	**Theory of mind network** / regions of the brain more active than baseline when humans think about other humans (Fletcher et al., 1995; Saxe & Kanwisher, 2003; Mitchell, 2004, 2005, 2009; Satterthwaite et al., 2018)
Specific knowledge about familiar people e.g. I am thinking about my brother specifically.	**Person-specific patterns** of regions activating **in the brain** (Gobbini et al., 2004; Patterson et al., 2007; Heisz et al., 2012; Thornton and Mitchell, 2017)
Self-construct e.g. I am thinking about myself, who I believe myself to be, and how I would act in a particular known or novel situation.	**Regions in the brain** activating in response to **thinking about oneself** as opposed to other people in the brain (Mitchell et al., 2005; Powell et al., 2009; Ma and Han, 2011)
Directional, third-person social interactions e.g. I am thinking about what my friend would think about my brother if he were to spoil the birthday surprise she was organizing for me.	**Neural activation patterns**, "representations", specialized for recognizing **agent and patient in social situations** (other research suggests that there may be an architecture for social cognition (Barrett and Satpute, 2013; Medaglia et al., 2015), that the human brain recognizes and remembers other people in social situations (Quadflieg and Koldewyn, 2017; Chen et al., 2017) and that studying other, especially familiar people is helpful for the study of social interactions (Redcay & Schilbach, 2019; Beckes et al., 2013)).
Directional, personal social interactions e.g. I am thinking about what my friend would think about me if I introduced her to my other friends in a social group she was unfamiliar with.	**Neural activation patterns**, "representations", specialized for recognizing **oneself as agent and patient** when processing information about a social interaction (other research suggests that neural representations of self may be multi-layered and more complex than those of other people (Newen & Vogeley, 2003), that studying self-referential representations is critical for interpreting the functions of the ToM (Apperly, 2008), that neural processing of other people's minds is impacted by self-referential activity (Ames et al., 2008), and that social interactions are fundamentally transformed by one's own participation in them (Schilbach et al., 2013))

Figure 1. Relationships between the scales of the objects of study, specifically the connections between social cognition in societal interactions and respective responses in the brain. The sections in blue indicate hypotheses studied in this research, with previous research supporting the current hypothetical approach.

Research questions

1) Are there distinct regions for **"agent" and "patient" neural representations** in the theory of mind network?
2) Are there distinct regions that neurally represent oneself as an "agent" and "patient" in the theory of mind network? If syntactic processing can be identified in cognitive processes, does it differ for the processing of **social information about the self versus others**?
3) Could we infer a **combinatorial architecture** in the theory of mind network? If so, to what extent does it follow the predictions of the Language of Thought hypothesis? To what extent is it generalizable for varied semantic content (self vs. other)?

Methods

Participants: 20 healthy, right-handed, native English speakers (3M, 16F, 1O; 18-29 years).
Stimuli: personalized stories involving hypothetical interactions between, or with, 3 personal acquaintances (first names only) involving the inferred emotions of "annoyed" or "grateful".
Scanning: 3T Siemens Prisma MRI scanner, 90 minutes total (75 mins active): 5 runs of the *other people task*, 3 runs of *self-others task*, 2 runs of the ToM localizer task (Saxelab, 2019).

Figure 4. Schematic representation of the blocked designs used in the *other people* task (left) and the *self-others* task (right), with an example of a run, and a condition block within it.

Representational Similarity Analysis (RSA)

Neuroimaging data was analyzed using RSA to transform a theoretical model of brain activity into a representational space, creating a simultaneously data- and hypothesis-driven analytical framework. Representational dissimilarity matrices (RDMs), shown below, demonstrate the inverse of the hypothetical correlation between two hypothetical conditions (Nili et al., 2014). Experimental data is then compared to these hypothetical models; the more similar empirical data is to a hypothetical matrix's prediction, the more closely it is assumed to 'represent' neural information. We used a focused searchlight analysis limited to the theory of mind network (Dufour et al, 2013) and searched for model similarities against baseline neural activity.

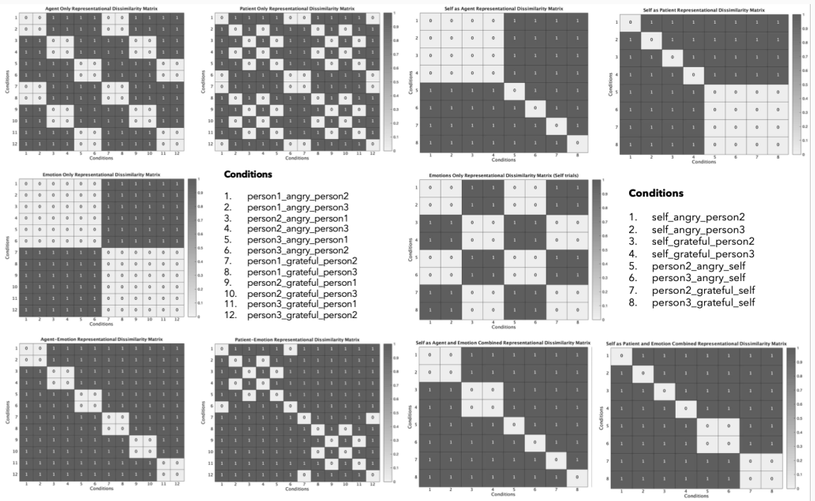

Figure 5. Representational dissimilarity matrices (RDMs) with modeled theoretical hypotheses for fMRI data analysis of the *other people task*, where 1 represents greatest hypothetical dissimilarity between two conditions. Hypothetical RDMs for the agent sub-region, patient sub-region, emotion sub-region, agent-emotion combined sub-region, and patient-emotion combined sub-region within theory of mind regions.

Figure 6. Representational dissimilarity matrices (RDMs) with modeled theoretical hypotheses for fMRI data analysis of the *self-others task*, where 1 represents greatest hypothetical dissimilarity between two conditions. Hypothetical RDMs for the self-as-agent sub-region, self-as-patient sub-region, emotion sub-region, self-as-agent-emotion combined sub-region, (e) self-as-patient-emotion combined sub-region within theory of mind regions.

Results

In preliminary analyses of the *other people task*, a statistically significant number of voxels within the mentalizing network resembled theoretical RDM models for emotion, agent-emotion and patient-emotion conditions; for the *self-others task*, the patient condition was significant, in addition to all those significant in the *other people task*. Significant voxels were identified across the mentalizing network (rTPJ, lTPJ, PC, dmPFC, mmPFC, vmPFC). Composite representations (role + thought) performed more significantly than single representations, suggesting emergent properties of combinatorial representations. A sample image of model similarity comparison is shown below. Further data (30 total participants) needs to be collected and analyzed before final conclusions can be drawn.

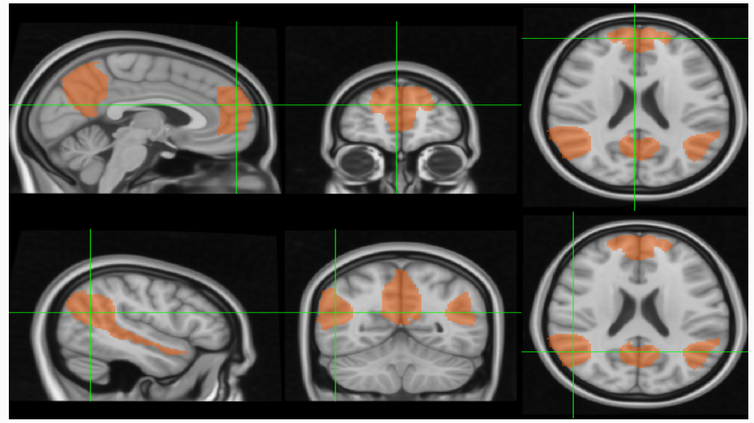

Figure 15. Regions in the brain, where the 'self as patient – emotion' hypothesis model RDM performed significantly above baseline neural activity in the *self-others* task. Identified using searchlight comparisons of modeled hypothetical RDMs with baseline (fixation) fMRI data for the 'self as patient - emotion' binding in the *self-others task*, where the model RDM matched at a statistically significant rate, when the False Discovery Rate was corrected for $p < 0.05$. The highlighted regions include the PC, rTPJ, lTPJ, rSTS, and dmPFC

Discussion

This thesis aims to empirically investigate *how* brain regions, consistently involved in social cognitive tasks, such as inferring the minds, thoughts and motivations of other people, perform their activity. Inspired by approaches from philosophy of mind (J. Fodor, H. Putnam, N. Chomsky), we implement an empirical investigation of the Language of Thought hypothesis (LoTH), which claims that cognition in the brain may operate similarly to the linguistic function of language: by representing mental properties and combining them according to "syntactic" rules of use within a symbolic system.
While definitive conclusions remain premature, we observe some emergent properties linked to our hypothetical models of the representations of semantic categories, distributed across the theory of mind network and are continuing data collection, analyses and interpretations.
From a linguistically pragmatic perspective, I suggest that confirming the indexical relation between social cognition and neural representations may be more challenging than traditionally assumed in cognitive neuroscience, and that the design of the current study is not well equipped to test the *symbolic* nature of brain function, and thus the *symbolic* claims of the LoTH. I also argue that future iterations of studying cognitive architecture should avoid monolinguistic biases in order to prevent linguistic-cognitive confounds. I further suggest that carefully designed cognitive empirical studies hold the promise to demonstrate metapragmatic properties of thought.
Through a socio-anthropological lens, I examine the implications of studying social cognition in non-social settings, such as those of neuroimaging, and evaluate the importance of positionality and ideology in the conduct of brain-mind research.
This research hopes to stimulate interdisciplinary interest and discourse on the architectural organization of cognition, and its potential implications for social relations.