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

Emilé Radyté | Supervised by Mitchell (Psychology) & Harrass (Anthropology)

Abstract
Humans navigate social environments daily: understanding the minds and thoughts of others, known as mentalizing, is crucial to performing this activity. Recent advances 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-other tasks) or do not include themselves (other people tasks). 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.

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, 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, 10: 18-29 years).
Stimuli: personalized stories involving hypothetical interactions (HIs), stories involving 3-person acquaintances (first names only) involving the inferred emotions of “annoyed” or “grateful”. Scanning: 31 Siemens Prisma MRI scanner, 90 minutes total (75 mins active): 5 runs of the other person task, 3 runs of self-other task, 2 runs of the ToM localizer task (Saxelab, 2019).

Figure 1. Representational Similarity Analysis (RSA) Imaging neuroactivity 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 most 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 across baseline activity.

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 most 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 across baseline activity.

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 emerging 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 monolingual 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 neuromaging, and evaluate the importance of postdiction 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.

Figure 15. Regions in the brain, where the ‘self as patient−emotion’ hypothesis model RDM performed significantly above baseline neural activity in the self-other task. Images show no other significant differences. The significance was based on a threshold of p < 0.05.