Lecture notes - social cognition part I

Ivana Konvalinka

October 28, 2025

1 Introduction

Social cognition is the sum of all processes that allow individuals to engage in successful interaction with one another [4]. It constitutes cognitive mechanisms underlying social behaviour and the processing of social information. Why is this important? Our brains are shaped through interactions with other people - i.e., how we think, feel, perceive, move, make decisions, remember, etc. are all influenced by the social interactions we have experienced in life. From the moment we are born, we are exposed to social interaction, e.g., observing and interacting with our parents and older siblings - and we learn about the world through these interactions.

But what is it that makes humans uniquely social, more than other animals?

For one, humans have language - we can express our thoughts, ideas, goals, and emotions to each other in a complex manner using language. Unlike other animals, humans have created culture. This allows us to learn from people who are no longer alive. Humans can also consciously reflect on themselves and others - we have metacognition, which is awareness of one's own thought processes. Metacognition allows us to have control (or delusion of control, but we will not open this can of worms) over many of our social mechanisms.

In addition, the human brain (the neocortex) is proportionally much larger in humans compared to other primates/mammals. The neocortex is the outermost layer, the largest part of the cerebral cortex. It comprises many brain areas involved in social cognition, such as those engaged in processing of language, empathy, emotion regulation, and thinking about others. Robin Dunbar [2] proposed the social brain hypothesis, which posits that there is a correlation between social group size among primates and their relative neocortex volume. The larger the relative neocortex volume (relative to body size), the bigger the social group size within which the species belongs, and interacts with. Therefore, we are hardwired to interact with other people.

Why is all this relevant for AI? Many industries such as healthcare, transportation, military, etc., are increasingly beginning to implement robots and virtual assistants that are imbued with a social artificial intelligence. Healthcare, in particular, may have have a need for social robots as there is a rising proportion of people aged >65, which is placing a strain on healthcare systems. For example, social robots are being implemented for healthcare for the elderly and stroke recovery - assisting with physical tasks, cognitive issues, health management, and companionship. AI has also been useful in evaluating and determining treatment in orthopedic surgeries, or driving marketing decisions in business.

But these implementations are often deterministic, hence developed as a system of rules that does not adapt well to new circumstances or environments. Importantly, the current state-of-the-art AI performs poorly in understanding other people's needs and intentions, learning about other people through experience, as well as being able to predict what other people are going to do next - which are the pinnacles of human social cognition. If we take autonomous vehicles as an example, a smart car needs to be able to make reliable predictions about human behaviour in real time. What if a child suddenly decided to cross the road in front of a car? One may argue that deep learning neural networks can identify human actions well in videos and in motion patterns - but they cannot predict human behaviour in real-time the way humans can, and adapt accordingly.

Below, I will first introduce some of the mechanisms that are crucial to social interactions, with a focus mostly on mentalizing/Theory of Mind. I will briefly consider how these are implemented in AI, which you can read about further if interested. Then, we will discuss how these are measured/quantified in the field of social cognition. Finally, I will discuss how we can apply AI to better understand social cognitive mechanisms. Therefore, the role of AI is two-fold: we can consider how understanding social cognition can help us design better artificial social agents; and second, we can use AI to better understand the underlying social cognitive mechanisms. These are two very differnet applications.

2 Social cognitive mechanisms

The field of social cognition is thus devoted to the study of social cognitive mechanisms. These mechanisms can be behavioural (e.g., we mirror other people's behaviour, and learn through mirroring), or neural (e.g., the brain areas or processes that are engaged during social interaction).

Many social processes are automatic, in other words, engaged without awareness. For example, we make assumptions about other people - how trustworthy they are, their emotions, their intentions - automatically, based on their appearance. In fact, we make an assumption of people based on the first 100 ms of seeing their face [4]. We also follow other people's gaze in order to know what they are focusing their attention on. Many of these mechanisms are related to social learning - we learn about the world and about others and ourselves from gaze following, mirroring, social referencing, imitation, etc. We will not discuss these mechanisms at length here.

However, there are also higher-level processes engaged in social cognition that are not automatic. For example, after we have had an awkward interaction, we may reflect on it consciously, and think about why it was so awkward, and whether we perhaps misunderstood the other person's intentions. These high-level processes increase possibilities for group actions, and allow for uniquely human communication to emerge. They also allow us to reflect, and change our subsequent social behaviour.

2.1 Theory of Mind

Mental states, such as desires, goals, or intentions, are invisible. But we can learn about and infer other people's mental states by observing their movements, direction of gaze, and other behaviours. We also learn about them by getting to know people - thus, we are better at understanding the mental states of our close family members, than of strangers. For example, if we see someone focusing their gaze intensely with eyes wide open, we may infer that they just spotted something they are afraid of. Or, if we're sitting at home, and our brother makes a similar face, but there's a subtlety in his face that we recognize, from which we may infer that he's intentionally teasing/deceiving us (pretending to have spotted a big spider), which is one of his usual pranks. We can also infer more subtle mental states, such as confusion or indifference, from the way people talk, move, signal with their eyes or hands. This ability has a crucial role in allowing us to predict others' future behaviour, and therefore, successfully interact with others.

The process of attributing mental states to others is known as mentalizing. Mentalizing thus refers to "reading others' minds". The ability to predict and explain other people's behaviour in terms of mental states is known as having a Theory of Mind. For example, imagine you have entered the elevator, and the doors are about to close. But you see someone frantically running towards the elevator. In the absence of verbal communication, from their facial expression and pace, your Theory of Mind allows you to anticipate that they are trying to reach the elevator before the doors close, as they need to get to another floor as fast as possible. Thus, it allows you to anticipate the needs and actions of others. As a nice person with such an ability, you would then hold the door open for them, so they could make the elevator in time.

We all have this ability as healthy adults, though perhaps to different degrees. It is still debatable when Theory of Mind emerges in development. The traditional belief in infant cognition was that Theory of Mind emerges from the age of 4, and that infants are unable to understand others' beliefs and intentions, or to imagine others' experiences. This is because children before the age of 4 are unable to pass a famous psychological test, called the Sally-Anne test, which measures a person's ability to attribute false beliefs to others. You can read about the test here: https://en.wikipedia.org/wiki/Sally-Anne_test. More recent theories consider that Theory of Mind can be both explicit or implicit. Older children can pass this task when explicitly asked to reflect about someone's mental states. But implicit mentalizing may develop earlier, and infants younger than 2 years of age may already begin to spontaneously attribute mental states to others. For example, these studies use eye tracking to measure whether infants look longer at certain events - those that violate their expectations

- i.e. regarding other people's beliefs. Implicit abilities thus do not require the need for language or executive processing. This is still very much debated.

Children with autism typically cannot pass the Sally-Anne test even up to their teens. This test was used some decades ago to better characterize the social cognition deficits in people with autism, showing impairments in Theory of Mind. However, this is a little more controversial today, as children with autism may also have language deficits, which may be a confound. And some are able to pass the test.

But what does it take to posses a Theory of Mind? Could we give AI a Theory of Mind? This is a very difficult task. One sophisticated AI (ToMnet) can solve tasks such as the Sally-Anne test [5]. But it would fail when faced with an unknown dynamic environment or person. You can read more about ToMnet here: https://singularityhub.com/2018/09/19/thinking-like-a-human-what-it-means-to-give-ai-a-theory-of-mind/#sm.0000w1jyxo1xaf4ywtv1ez8977dfv.

Large language models have recently surpassed previous AI models when it comes to performance on Theory of Mind tasks. Across a series of tests, GPT-4 models were shown to perform at, or sometimes above, human levels at identifying indirect requests and false beliefs, but struggled with detecting "faux pas". But does this mean that they truly understand? ToM tests such as the Sally-Anne test may be reliable proxies for assessing general abilities in humans, but not in AI systems. AI and cognitive science researchers have thus proposed that we need new scientific methods to measure mechanisms of understanding.

Importantly, apart from using tests such as the Sally-Anne test, which are easy to pass for adults, how can we measure Theory of Mind in humans (or artificial agents), or other social cognitive mechanisms? The following section addresses this.

2.2 Measuring and quantifying social cognitive mechanisms

There are two main approaches to studying social cognition (arguably more, but let's say 2 for the sake of simplicity). One is by studying individuals responding to social stimuli and measuring their behaviour and brain activity (individual approach); the second, more recent, is by studying people engaged in interaction, while measuring their beahviour and brain activity (joint action approach).

Until a decade and a half ago, the field of social cognition/neuroscience only employed the individual approach - focusing on individuals isolated from a real social interaction. This was for simplicity reasons, as measuring brain activity using functional magnetic resonance imaging (fMRI) was only feasible with one person in the scanner. This was also more feasible, as it was easier to control for the stimuli the participants were presented with, and to isolate the social aspect of the person's response. This meant, however, that people interacted with computerized responses, abstract or artificial stimuli, and everything we understood about social cognition until recently was measured in the absence of real social interactions with other people. Therefore, this approach was challenged, and it was proposed that social cognition is fundamentally different when we engage in social interaction - which gave rise to the joint action approach.

Here, we focus mostly on the first approach, though I will briefly introduce the idea behind the second approach at the end.

2.2.1 Individual approach and the social brain

Many of the tasks designed to study social mechanisms include mirroring (e.g., watching a hand performing goal- or non-goal directed movements), imitation, or mentalizing tasks - many of them from a purely observer perspective - as well as appropriate control conditions (e.g., non-social conditions, such as observing houses instead of faces, or watching a tool move the way the hand did). While the participants observed or engaged in such tasks, their brain activity and responses (e.g. in a social decision making game) were recorded. Using fMRI or electroencephalography (EEG), among other modalities, many of the studies thus contrasted the brain areas (or brain oscillations in the case of EEG) that were "activated" (this is in relation to the BOLD signal in fMRI, which is not exactly a brain activation, but this is outside of the scope of this course) or modulated (in the case of oscillations) during the social task, with brain areas "activated" during the control condition. This would help isolate the social aspect. (Note: this is an oversimplification).

In 1990, Brothers proposed that there is a restricted set of regions dedicated to social cognition [1]. The evidence came from lesion studies in monkeys - i.e., lesions to the amygdala and the orbital frontal cortex altered monkeys' social behaviour, made them socially isolated, etc. This set of regions became known as the "social brain" - regions thought to be dedicated specifically to social functions.

Numerous fMRI studies corroborated this hypothesis, consistently finding activation in the same brain areas during the social tasks in contrast to non-social tasks, proposing a social mechanism for each of these brain regions. For example, the superior temporal sulcus (STS) is thought to be responsible for the processing of faces and biological motion, as it consistently activates to human biological motion in contrast to non-biological motion. The temporoparietal junction (TPJ) and the medial prefrontal cortex (mPFC) are thought to be involved in mentalizing, with the TPJ activating for immediate goals and desires, while the mPFC activates to more enduring traits of others. The amygdala is engaged in emotional processing. And so on. (Note: You do not have to remember the brain areas for the exam).

While there are many proposed processes of the social brain, Chris Frith proposed that the main function of the social brain is to make predictions during social interactions [3]. According to him, the brain has two problems to solve:

1. Read the mental state of the person one is interacting with, based on their actions.
2. Make predictions about future behaviour on the basis of that state.

Can you place this in a Bayesian framework? (We will do this in class.)

But how do we read other minds, despite their inaccessibility? In other words, how do we understand others' actions, emotions, and intentions?

There are two main theories. The first is called **Theory-Theory (TT)**. It postulates that we infer the mental states of others on the basis of our own mental states. Therefore, according to TT, we do not "experience" the mental state of others ourselves, but we infer them based on our experience. We thus form theories of others' mental states, and test them against each other.

The second theory is called **Simulation-Theory (ST)**. According to ST, we simulate the actions or beliefs of others by activating the corresponding representation of those actions or beliefs within ourselves. It thus postulates that we experience the mental states of others ourselves, understanding others' minds literally through ourselves.

2.2.2 Mirror neurons and the human mirror neuron system

Neuroscience provides some support of Simulation-Theory, through the discovery of mirror neurons. Mirror neurons are neurons that have been found in the brains of macaque monkeys - in ventral premotor cortex area F5 and the inferior parietal cortex - which fire both when the monkey performs a motor action (during action execution), and when the monkey observes another individual performing the same or similar action (during action observation). These neurons thus support the link between perception (action observation) and action (action execution). But their function has been long debated (though this debate has been more silent in recent years), and their significance is still controversial.

In humans, we cannot test for direct existence of mirror neurons, as this would require cutting into the human brain. Therefore, the existence of mirror neurons in humans is still very much debated. There is some evidence found with fMRI and EEG. fMRI studies have shown that areas that correspond to locations where the monkey mirror neurons were found are "activated" when we produce an action, and when we observe someone else producing a similar action. This became known as the human mirror neuron system.

With EEG, when we produce an action, the amplitude of 10 and 20 Hz oscillations over the motor cortex (known as the mu rhythm) is suppressed (decreased). This is in contrast to the rest position, when the mu rhythm amplitude is high. The mu rhythm is thus thought to be involved in the control of voluntary movements, and the suppression of its amplitude is thought to represent release from inhibition (or excitation) of the motor cortex, allowing us to move. Interestingly, the mu rhythm also suppresses (to a lesser extent) when we observe someone producing the same or similar movements. There is a lot of disagreement on whether this is related to the human mirror neuron system. In any case, it seems to be coding for both action and perception.

More broadly, many other human mirroring systems have been observed. For example, the same brain areas are activated when we experience pain and see someone receiving a painful stimulus; somatosensory areas (i.e. areas that process touch, pressure, temperature, etc.) are activated when we see someone else being touched, etc. The location of these "mirroring systems" depends on what is being observed.

2.2.3 Measuring neural mechanisms underlying social interaction

So far, we have discussed the social brain, and mechanisms that are responsible for the processing of social information. But how do these differ from mechanisms that are needed for us to engage in real social interaction with others? This is something we will focus on next week.

For now, let's consider that we know very little about the brain processes that allow us to engage in successful social interaction. We want to measure these, and thus we design a simple interactive experiment. For example, we ask pairs of people to finger tap together, and synchronize their taps. We use this very simple task to address how people coordinate their actions in real-time, by integrating perceptual information (taps from another person) with their own actions. We now have an action-perception loop, where the perceptual information is not just from our environment, but from another person. This person is also noisy, but in addition, they have their own actions and mental states, as well as predictions about our actions and our mental states. We now have a bidirectional action-perception loop (one person's action output is another's perceptual input, and vice-versa).

We also need a control condition, to contrast the interactive mechanisms with another person with either non-interactive mechanisms (tapping alone), or interaction with an AI/computer. Let's introduce a very simple control, which is a completely predictable computer metronome that people have to synchronize to.

Now our experiment looks something like this. Pairs of participants are asked to synchronize their taps a) together with another person (who is unpredictable, but adaptable), or b) with a computer metronome (which is predictable, but not-adaptable) - a control for our social condition. We want to measure the neural mechanisms underlying this simple social interaction. We measure EEG from both participants engaged in interaction. We decide to focus on 10 Hz oscillations, which we know are involved in producing movements, as well as perceiving movements, when they originate from the motor cortex. But widespread 10 Hz oscillations (known as the alpha rhythm, around 8-12 Hz, often centered at 10 Hz when not originating from the motor cortex) are also involved in attention, and may be differently involved in a social versus a non-social task.

What do we do with the two-person EEG data to make sense of it? We will do this exercise in class. Specifically, you will get to classify the preprocessed EEG data - 10 Hz oscillations - across the scalp into interaction with human vs. interaction with computer conditions. We will use a logistic regression to classify the two conditions, based on the brain data. We will use forward sequential feature selection to select the best features. See lecture notes on cross-validation, forward feature selection, and logistic regression in the Exercise folder.

References

[1] L. Brothers. The social brain: a project for integrating primate behaviour and neurophysiology in a new domain. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, 1:27–51, 1990.

[2] R. I. Dunbar. The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, pages 178–190, 1998.

[3] C. D. Frith. The social brain? *Philosophical Transactions of the Royal Society B*, 362:671–678, 2007.

[4] C. D. Frith and U. Frith. Social cognition in humans. *Current Biology*, 17(16):R724–R732, 2007.

[5] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. A. Eslami, and M. Botvinick. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227, 2018.