Theory Theory as an Ideal Theory of Mind

Humans are social animals, and as such understanding each other's thoughts, feelings, and beliefs is a crucial aspect of our interactions. This ability, broadly known as Theory of Mind, allows us to interpret and predict others' behaviors. One prominent explanation for how we develop this theory is Theory-theory. Gopnik and Wellman (1992) explain the fundamental idea of Theory-theory through children in their paper, saying that their "early understanding of mind is an implicit theory analogous to scientific theories, and changes in that understanding may be understood as theory changes."(Gopnik and Wellman 146). This "implicit theory" is an intuitive understanding of how the mind may work, and as we observe people and gather more evidence we refine this implicit theory to make it more accurate. For instance a child may initially believe that people dislike rain because it makes them wet. However as this child ages they may observe people liking rain on cozy days, when sad, or when dancing in it. These can all be used to refine the child's theory and better predict how people will feel or act when it rains.

Another competing explanation is called Simulation Theory, which suggests that we understand others minds' by mentally simulating their experience. Stitch and Nichols say "The Simulation Theory as I present it holds that we explain and predict behavior not by applying a theory but simply by exercising a skill that has two components: the capacity for practical reasoning … and the capacity to introduce 'pretend' facts and values into one's decision-making." (Stitch and Nichols 5). Instead of forming abstract theories, we "put ourselves in their shoes" to infer emotions and thoughts. For example, upon seeing a friend get rejected by a school, they may imagine themself getting rejected, understand the feeling of sadness, insecurity, and worry, and project that onto their friends mind.

While both of these theories aim to understand humans' theory of mind through different avenues, they share key similarities and differences that provide deeper insight into how we interpret other minds. In this paper I will explore these similarities and differences, and argue that Theory-theory offers a stronger explanation due to a more universal foundation, while addressing a counterargument.

Two key similarities between simulation theory and Theory-theory are that they both offer a developmental perspective and both assume an innate foundation for understanding minds. Theory-theory explicitly frames development as a process of refining a mental theory over time based on experience. Gopnik and Wellman suggest this to be a gradual change, explaining that "there is a change from one mentalistic psychological theory to another somewhere between 2.2 and around 4.2. The change is not a simple all-or-none one, but rather involves a more gradual transition from one view of the mind to another" (Gopnik and Wellman 146). This suggests that children's understanding of minds evolves continuously as they accumulate more observational data and revise their implicit theories. Simulation also assumes development, referring back to the definition presented by Stich and Nichols, the idea of simulation theory as "exercising a skill" also builds around the idea that it can be developed over time. In this theory, the cognitive skills of simulation and practical reasoning strengthen through repeated use. The second similarity between the two is their shared assumption of an innate ability to understand the mind. As Gopnik and Wellman describe, a "child's early understanding of mind is an implicit theory" (Gopnik and Wellman 146). This suggests that children start with a foundational, though possibly incorrect, theory of other minds that is refined over time. Stich and Nichols reinforce this notion from the perspective of Simulation Theory, stating that "Gordon later suggests that the capacity to simulate in this way may be largely innate: '[Evidence]

suggests that the readiness for simulation is a prepackaged "module" called upon automatically in the perception of other human beings'" (Stich and Nichols 5). This suggests that the ability to simulate is not something entirely learned but rather an inherent cognitive function that individuals draw upon. Ultimately, while each theory differs in its mechanisms, both recognize that humans are born with an initial ability to interpret other minds, and both emphasize that this ability develops and refines with experience.

Despite their similarities, the two theories also share significant differences. The mechanism through which they understand, how they develop, and our awareness of them differ. The key difference between the two is mechanism. Theory-theory defines a mechanism where we understand others by applying a set of psychological rules or principles, which we use to predict how others act. The mechanism for Theory-theory develops a folk psychology to judge from, much like a scientist refining a theory. In contrast, Simulation Theory's mechanism does not utilize rules but rather uses our own mental and cognitive processes as a model for understanding others. The mechanism for Simulation Theory relies on our own mind to judge from, rather than an external set of principles. Another important difference is the developmental process of the two. Theory-theory develops by learning rules through experience. The gradual change Gopnik describes results from observing others and revising mental hypotheses over time. Simulation Theory, however, develops through improving simulation skills. A child may begin with a poor ability to simulate others, but with more emotional and social experience, their skill becomes more accurate, resulting in a developmental process that is more akin to strengthening a skill than learning rules. The final difference between the two is how aware we are of the process. The paradigm of Theory-theory suggests we have our own folk psychology and actively apply these rules to situations. Gopnik explains that "in some cases, the theory may

be partly accessible to consciousness; the agent can tell us some of the rules or principles he is using" (Gopnik and Wellman 146). The ability to explain the rules one applies suggests a more consciously available process. However, Simulation Theory's framework of simulating another's experience differs. Stich and Nichols explain that "this whole process may be largely unconscious. It may be that all you are aware of is the prediction itself. Alternatively, if you consciously imagine what the target of your prediction will do, it could well be the case that your imagination is guided by this simulation rather than by some internally represented psychological theory" (Stich and Nichols 4). This suggests that simulations can operate beneath conscious awareness, meaning that while we arrive at an answer, we may lack the ability to explain our reasoning or the origin of our prediction. Ultimately, while both theories aim to explain how we understand others, they differ fundamentally in their cognitive mechanisms, how they develop over time, and whether we are conscious of their operation. These distinctions shape how each theory accounts for our ability to interpret and predict human behavior.

I argue that Theory-theory is a stronger theory because of its universal basis. By universal basis I mean that the innate skills that are assumed under Theory-theory can truly be held by everyone, whereas those assumed for simulation theory are not. I will frame my argument for Theory-theory as an argument to the best explanation. With the premises as follows:

1. People with autism spectrum disorder (ASD) and antisocial personality disorder (ASPD) struggle with intuitive social understanding for respective reasons.
2. These people's disorders prevent them from having the same innate simulation skills as neurotypical people.
3. These disorders do not disrupt their ability to hold rules differently from neurotypical people.

4. Because people with ASD and ASPD still interact socially, the rule based framework for Theory-theory is the best explanation for theory of mind.

I will now add a more detailed explanation to my premises before handling a counterargument. People with ASD struggle with social interaction, and understanding others in the ways neurotypicals do. They find it difficult to put themselves in others' shoes, and their social understanding often takes a more rules based shape. For example "if someone frowns, they're upset" or "If they look away, they may be bored" are common rules for people with ASD. The struggles that come with this disorder make the possibility of simulating unlikely, however rules still remain productive. People with ASPD struggle with empathy, whereas a neurotypical may feel discomfort when a friend cries, a person with ASPD may not because they lack a connection to the other person's feelings. This presents a tremendous roadblock for simulation theory, as the lack of empathy makes simulating another's experience very difficult. However, both people with ASD and ASPD socialize, and with a success befitting of understanding others' minds. Although they both face their own difficulties, people with ASD learn to mask, hiding autistic traits to fit norms, and people with ASPD have historical examples of being charismatic, Ted Bundy being a notable one. Thus because those with ASD and ASPD cannot hold the same automatic simulation of neurotypicals, but all three groups can hold rules the rules required under Theory-theory, Theory-theory's requirements are universal whereas the "module" for simulation theory is not.

A counterargument to this is that although people with ASD and ASPD may not be able to access the unconscious simulation Stich and Nichols describe, their dispositions do not bar them from accessing simulation deliberately and with effortful analysis. For example someone with ASD may not automatically feel sadness upon another's sadness, but if they recall their own

experience and walk through it they could arrive at an approximation. Furthermore, someone with ASPD may recall their own experience being sad, and successfully predict a sad person's behavior.

Although this counterargument is clever, it ultimately serves to strengthen my argument rather than tear it down. The process described above would still require a rule adjacent framework as a consequence of its deliberate simulation. Because the process does not occur automatically like it does for neurotypicals, the structure of a simulation would be rule based. For example, upon seeing a friend cry, both someone with ASD and ASPD would have to walk through a plan. In the simplest sense it would be structured as:

1. My friend is crying, I don't understand this.
2. I should understand it (whether through care, social pressure, of personal gain)
3. What is a similar experience I can remember to understand?

I argue that this process itself is an example of Theory-theory, the people with ASD and ASPD are walking step by step through a framework and applying a rule, to call upon a memory, in order to arrive at a theory of what their friend is thinking and feeling. The framework required for this counterargument to work, would still require a rule based system and thus acts as another representation of Theory-theory instead of an edited version of Simulation Theory.

In conclusion, Theory-theory provides a more universal and flexible framework for Theory of Mind. While Simulation Theory assumes an innate simulation ability that not all individuals possess, Theory-theory allows for a learned, rule-based approach that applies across different cognitive profiles, including ASD and ASPD individuals. Even in cases where ASD and ASPD individuals engage in effortful simulation, their process remains structured and

rule-driven, reinforcing the principles of Theory-theory rather than Simulation Theory. Thus, Theory-theory is the best explanation for how humans understand the minds of others.