

---

# Framework for a multi-dimensional test of theory of mind for humans and AI systems

---

**Caoimhe Harrington Stack**

**Sarah Myers**

**Effat Farhana**

**Aviv Roskes**

**Xinyu Shen**

**Simeng Zhao**

**Angela Maliakal**

**Roxanne Rashedi**

**Joel Michelson**

**Maithilee Kunda**

Computer Science, Vanderbilt University, Nashville, TN

C.STACK@VANDERBILT.EDU

SARAH.A.MYERS@VANDERBILT.EDU

EFFAT.FARHANA@VANDERBILT.EDU

AVIV.D.ROSQUES@VANDERBILT.EDU

XINYU.SHEN@VANDERBILT.EDU

ELLEN.SIMENGZHAO@GMAIL.COM

ANGELA.MALIAKAL@TUFTS.EDU

RRASHEDI96@GMAIL.COM

JOEL.P.MICHELSON@VANDERBILT.EDU

MKUNDA@VANDERBILT.EDU

## Abstract

We propose a new framework for a multi-dimensional test of theory of mind based on scene comprehension, i.e., watching a short movie clip and then answering questions about characters' emotions, beliefs, goals, and values. Unlike other tests of theory of mind that have undifferentiated questions or questions of just a few types, our test framework is based on a new three-component model that we developed based on our research on social cognition in autism. Our framework divides theory of mind into three primary domains of perception, knowledge, and reasoning, and 19 subdomains. We expect that a test built using this framework would be valuable for understanding human learning and cognition, including in neurodiverse populations such as individuals on the autism spectrum, and also including informing the design of computational models of ToM learning and reasoning. In addition, we expect that this test will serve as a more challenging and interpretable benchmark for AI systems designed for theory of mind reasoning.

## 1. Introduction

Theory of Mind (ToM) is the ability to infer about and understand others' beliefs, desires, intentions, goals, and values. It underpins social interaction as it allows one to represent what another person is thinking or feeling, figure out why they are feeling that way, and predict how they might respond to future situations. These ToM abilities have long-term impacts on one's well-being and social skills.

Strong ToM skills are associated with positive, sustained relationships as well as overall social strengths (Fink et al., 2015; Liddle & Nettle, 2006; Peterson & Siegal, 2002). Meanwhile, reduced ToM skills are associated with many negative outcomes. For example, ToM is thought to play a critical role in the development of social skills in individuals with Autism Spectrum Disorder (ASD) with many individuals with ASD experiencing difficulty with ToM skills (Baron-Cohen,

2000; Kimhi, 2014; Peterson et al., 2016). It has been found that adolescents with ASD have stronger feelings of loneliness than their neurotypical peers (Bauminger et al., 2003), as well as weaker friendships (Bauminger & Kasari, 2000). In adults with ASD, it has been found that feelings of isolation and stress about initiating social interactions are common (Müller et al., 2008). Difficulties aren't limited to just the social realm, as children with ASD with social difficulties have been found to have greater mental health difficulties (Ratcliffe et al., 2015).

In addition, ToM has been shown to be related to many areas of academic and intellectual development. For instance, a child's knowledge of mental states helps that child to develop positive relationships with their teachers (Garner & Waajid, 2008). In turn, these positive relationships with teachers can contribute to academic achievement (Hamre & Pianta, 2006). ToM also includes the ability to explain others' behavior through cause and effect explanations of their mental states (Wellman & Lagattuta, 2004). Engaging in explanations such as these are important for learning in a variety of settings, including school (Wellman & Lagattuta, 2004). For example, it was found that preschoolers' emotional expressiveness, regulatory abilities, and knowledge significantly predicted preliteracy performance (Curby et al., 2015). In addition, research has found significant correlations between social-emotional skills in kindergarten and adult outcomes in areas such as education, employment, and mental health (Jones et al., 2015).

Though there are many clinical groups which show reduced ToM, such as those with depression (Bora & Berk, 2016), those with schizophrenia (Sprong et al., 2007), and those with ASD (Kimhi, 2014), neurotypical humans develop ToM skills through informal, natural social interactions with people around them, a process also referred to as '*developing commonsense psychology*' (Moore, 2013). However, there is much that is still unknown about how humans learn ToM, though several important ingredients have been identified, for instance: (1) ToM is related to many socially relevant precursor skills such as joint attention and gaze following (Tomasello, 2009). (2) ToM also seems to require meta-representational capabilities, i.e., to represent counter-factuals, pretending, and other complex, multi-layered belief states (Leslie, 1987). (3) ToM is strongly related to language ability, even in studies that controlled for language requirements of ToM tasks (De Villiers, 2007).

On the AI side, while there has been some work in developing computational theories and models of ToM (see Section 4 below), AI models of many kinds of ToM reasoning still lag far behind what humans (even young children) can do. There are at least two ways in which developing robust computational models of ToM learning and reasoning is important. First, incorporating ToM in machines will enable better human-AI collaboration. For example, an AI system designed to interact with elderly patients needs to understand human language, including both explicit and implicit references to emotions, beliefs, and other internal states of the patient. As ToM also involves predicting others' behavior, an AI trained effectively to deal with ToM may become better at predicting human behavior based on the emotion, beliefs, intentions, and values that it infers.

Second, computational models of learning can serve as testbeds for examining hypotheses about human learning and development, which can better inform research on atypical developmental trajectories, including methods for early screening and intervention. For example, there is a recent spate of work using deep learning models to study hypothesized mechanisms of early language learning in infants (e.g. Clerkin et al., 2017). We expect that computational models of ToM learning have the potential to serve a similar function.

**A problem in current ToM research.** Most current standardized ToM assessments (e.g., ToM Task Battery by Hutchins et al. (2008)) rely on simple stimuli whereas in real life ToM tasks are multimodal, complex, and conflicting. Fletcher-Watson et al. argued that ToM is composed of many subskills, and that, *“it is understood that different age and ability levels require support developing different sub-skills and mapping out these relationships would be of value”* (Fletcher-Watson et al. (2014), p. 23). However, because ToM is so broad and multi-faceted, it is difficult for a single assessment to cover all relevant skills and subskills.

Our proposed framework, the Multi-Dimensional-ToM (MD-ToM), fills this need by providing detailed scores not just for overall ToM but also across three primary domains of perception, knowledge, and reasoning, as well as numerous subdomains. We list our contributions as follows:

- A conceptual framework representing three dimensions of ToM, the MD-ToM.
- A construct map to design questions using our proposed MD-ToM and example questions.
- Implications of MD-ToM for research in AI and human learning.

## 2. Theory of Mind Tasks

Four commonly used ToM tests in human subjects research are discussed briefly below to highlight the range of skills assessed in ToM.

**The Sally-Anne Task and other false belief tasks.** False belief tasks are classic ToM tasks which test whether subjects can understand that another person can hold a false belief (Bosco et al., 2016). For instance, if Sally has a marble that she left in her basket, that Anne then hides in a box, where will Sally look for her marble (Baron-Cohen et al., 1985)? False belief tasks can be of the first order- "Where does Sally think her marble is?"- and of the second order - "Where does Anne think that Sally thinks her marble is?"- which requires understanding nested beliefs.

**Happé’s Strange Stories.** Participants read about social situations which, to be interpreted correctly, require an understanding of complex social reasoning such as pretenses or lies (Happé, 1994). Similar tests, such as the Faux Pas test, also measure children’s abilities to understand that a Faux Pas has occurred in social situations (Baron-Cohen et al., 1999).

**Reading the Mind in the Eyes.** In the Reading the Mind in the Eyes test, participants are shown static images of eyes displaying specific emotions and are tasked with matching the displayed emotion to the correct word, such as 'angry' or 'sad' (Baron-Cohen et al., 2001).

**Theory of Mind Assessment Scale (Th.om.a.s.).** The Th.om.a.s. is a semi-structured interview designed to assess ToM along four different scales that measures six facets of Theory of Mind. It assesses subjects along first and third person understanding, first and second order beliefs, and allocentric and egocentric understanding (Bosco et al., 2009).

The above list of assessments makes it clear that ToM covers a large range of social cognition. However, as pointed out in prior research (Fletcher-Watson et al., 2014), different ToM assessments target very different aspects of ToM. There is a need for assessments that cover a broad range of ToM skills, including both different kinds of skills as well as different difficulty levels. Our MD-ToM assessment design fulfills this need.

### 3. Scene Comprehension in Theory of Mind

We developed the MD-ToM design based on our current research to develop a ToM learning intervention for adolescents with ASD, including observational studies of how adolescents on the spectrum talk about social concepts in movie clips in comparison to neurotypical adolescents (Zi et al., 2020); how parents and caregivers of these adolescents viewed the social and ToM difficulties their children faced in everyday life (Rashedi et al., 2021); and a study of crowdsourcing ToM-related questions for movie clips (Chen et al., 2020). In developing the MD-ToM, we also draw from models of language and reading comprehension as analogies to ToM processing. We compare the scene comprehension we test in the MD-ToM to reading comprehension below.

Successful reading comprehension requires the merging of foundational skills, background knowledge, and reasoning in order to accurately infer a text’s meaning (Kintsch & Walter Kintsch, 1998). Reading begins with the ability to accurately read the words on a page, requiring word recognition and phonological awareness to decode difficult words. The next stage of reading requires integrating prior knowledge into current stage of comprehension. This prior knowledge can include both what has already been learned in the reading and general world knowledge that helps to link ideas to the text so that inferences can be pulled (Duke et al., 2011). Finally, reading requires reasoning about the text, drawing inferences about why something may have occurred, and predicting what may happen in the future (Duke et al., 2011).

Through MD-ToM, we propose that scene comprehension requires similar ToM skills. ToM builds off of foundational skills, learning how to interpret facial expressions and tone of voice for example. Then, background knowledge about schemata help one to bring relevant background knowledge to the scene and enrich one’s understanding of what is occurring. Finally, one must reason abductively about what has happened in the past to lead to the current state of affairs, and make predictions about what will happen in the future.

Importantly, just as a student being asked to read and comprehend a newspaper is tasked with real world comprehension, our scene comprehension is ecologically valid. That is, we propose that the MD-ToM can be used with clips from a wide range of sources displaying a wide range of social situations and these clips do not need to test subdomains in a vacuum. Rather, the MD-ToM is a framework that allows testing of naturalistic clips and thus naturalistic scene processing.

### 4. Related Work on ToM in AI

**Theory of Mind.** The evaluation of machine ToM has seen a growth in popularity. Works, including Rabinowitz et al. (2018), use a gridworld variant of a Sally-Anne task to test for second-order false beliefs. This same work presents a supervised learning model, ToMNet, to make explicit predictions about agents’ mental states. Google’s LaMDA large language model is shown to succeed at one question based on the Sally-Anne task (Thoppilan et al., 2022).

**Reasoning.** AI researchers have explored natural language inference (NLI), a form of deductive reasoning, in the context of machines. Authors Sap et al. released a knowledge graph, ATOMIC (Sap et al., 2019a), containing causes and effects of 24,000 events. ATOMIC contains six types of question: 1) Want, 2) Reactions, 3) Descriptions, 4) Motivations, 5) Needs, and 6) Effects.

**Social and Emotional Contexts.** The SocialIQ dataset (Sap et al., 2019b) contains 38,000 multiple choice questions (MCQs) on social and emotional contexts. The corpus contains contexts and reasoning questions about motivations, predictives, and emotional reactions.

**Visual Commonsense and Visual Question Answering.** A number of datasets, including CLEVR (Johnson et al., 2017), SHAPES (Andreas et al., 2016), and NLVR (Suhr et al., 2017), pose natural language questioning regarding synthetic images to test for primitive visuospatial skills like identifying attributes, counting, and basic spatial relationships. Another line of research is focusing on Social Visual Question answering (VQA) which requires AI systems to use human cognition-level visual understanding for images or video clips. Examples include Li et al. (2022); Zadeh et al. (2019); Zellers et al. (2019, 2022).

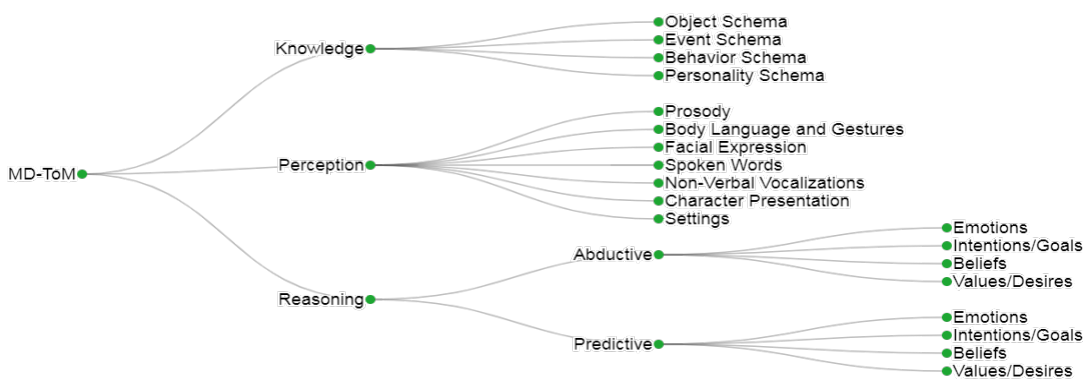


Figure 1. MD-ToM High Level Overview: 3-Component Model

## 5. New theoretical model of ToM:MD-ToM

### 5.1 MD-ToM Model

We propose a new Three-Component (3C) theoretical model of ToM which proposes that successful ToM requires the integration of various forms of (1) perception, (2) knowledge, and (3) reasoning, with further specialized subcategories of skills within each component. This model is most similar to the Two-Component model that proposes social-perceptual and social-cognitive components (Tager-Flusberg & Sullivan, 2000), but we further divide the cognitive component into knowledge and reasoning components. Figure 1 shows an overview of MD-ToM. We further identify 19 subdomains within these domains, where each subdomain identifies a distinct slice of ToM skills in terms of the cognitive content and types of inferences that are required.

### 5.2 Construct Map

We define three primary constructs (perception, knowledge, and reasoning) and their subdomains in terms of “construct maps” (Wilson, 2004) that describe the constructs and their subdomains. For each subdomain, we define three levels of understanding: **LOW**, **MEDIUM**, and **HIGH**.

These descriptions are specific to the ToM task of *scene comprehension* that we use for the MD-ToM, i.e., watching a film clip and interpreting ToM-related content for characters in the clip. All of the domains and subdomains listed for scene comprehension will have analogues in real-world ToM tasks, e.g., when perceiving and interpreting live social interactions, though of course real-world interactions are different in many ways as well.

We describe our MD-ToM domains and 19 subdomains below.

- **Perception.** Cognitive science researchers argue that social and cognitive components of ToM build upon social-perceptions of other people (i.e., making judgements about people’s mental state from their facial and body expressions) (Baron-Cohen, 1997; Hobson, 1993). Figure 3 presents our perception subdomain. We describe sub-domains of our MD-ToM perception subdomains below. These subdomains can be used in isolation or in combination to make inferences about a person’s emotional state, intentions/goals, beliefs, and/or values.
  1. Prosody: Prosody includes the suprasegmental elements of language, such as tone of voice, intonation, rhythm, stress, volume, which are a part of *how* words are said rather than *what* words are said (Cutler et al., 1997). Prosody can be used to understand the emotions of a person speaking (Luengo et al., 2005).
  2. Body Language and Gestures: How a person presents their body as an indication of how they are feeling or what they are trying to communicate (Phutela, 2015). Body language may include indicators such as where a person is facing or hunched shoulders. Gestures may include indicating signs or the hands moving to emphasize a point.
  3. Facial Expressions: How a person makes their face look to reflect what they are feeling or trying to communicate. Facial expressions can be used to accurately identify emotions (Kirouac & Doré, 1983).
  4. Spoken Words: The words that a person speaks. Speech can be literal or may be figurative, ironic, or otherwise misleading.
  5. Non-Verbal Vocalization: Non-verbal vocalizations include laughter, grunts, gasps, groans, and screams and can be used to communicate emotional expression and how a person feels (Sauter et al., 2010).
  6. Character Presentation: What a character looks like and is dressed like. Viewers may make inferences based on a character’s social group (Hammond & Cimpian, 2017) and may also use appearances to attribute qualities to an individual (Langlois, 1995).
  7. Settings: Identifying where a scene is taking place. Correctly identifying a setting allows viewers to make assumptions about the type of behavior that is expected in a location as well as what objects might be found in a location (Ramey et al., 2022).
- **Knowledge.** Figure 4 presents our knowledge subdomain. All subdomains of our knowledge construct rely on schemata, a collection of knowledge about an object, event, or concept (VandenBos, 2007). These collections of knowledge usually work to simplify one’s understanding of the world by providing a base set of assumptions about a situations. We propose the following four schemata for use in scene comprehension:

1. Object Schema: Knowledge about how objects operate in the setting one is in.
  2. Event Schema: Knowledge about how events should precede given the event taking place.
  3. Behaviour Schema: Knowledge about how people should behave given the setting.
  4. Personality Schema: Knowledge about how a person should act given their personality.
- **Reasoning (Abductive).** Figure 5 shows our abductive reasoning subdomains. Abductive reasoning is one of the three major types of inference in philosophy (Zalta et al., 1995). It requires inferring what has happened based on the conclusion you see. We propose that abductive reasoning is one of the necessary components of scene comprehension. That is, an observer needs to figure out given the observation of X, what is the most likely explanation for it. This requires observers to formulate ideas about what is driving an agent’s emotions, beliefs, goals, and/or values.
    1. Emotions: An emotion is a state that results in physical and psychological changes that influence thoughts and behaviors (Damasio, 2004). Given that a person X has emotion Y, why do they have that emotion?
    2. Intentions/Goals: An intention is an action in pursuit of a goal (Baldwin & Baird, 2001). Given that a person X has goal Y, why do they have that intention? (Phillips et al., 1998).
    3. Beliefs: Beliefs are ideas or principles which people judge to be true (Eccles et al., 2002). Given that a person X has belief Y, why do they have that belief?
    4. Values/Desires: Desires are emotionally charged motivations toward a targeted object, person, or activity that is associated with pleasure or relief from displeasure (Kavanagh et al., 2005). Given that character X has value/desire Y, why do they have that value/desire?
  - **Reasoning (Predictive).** Figure 5 presents our predictive reasoning subdomain. Prediction occurs when one makes an educated guess on what will happen next given the evidence already acquired. We propose that predictive reasoning is one of the necessary components of scene comprehension. That is, an observer needs to figure out given the observation of X, what will happen in the future. This requires observers to formulate ideas about what will occur based off of an emotion, goal, belief, or value that has already occurred.
    1. Emotions. Given that a person X has emotion Y, what will happen in the future?
    2. Intentions/Goals. Given that a person X has goal Y, what will happen in the future?
    3. Beliefs. Given that a person X has belief Y, what will happen in the future?
    4. Values/Desires. Given that a person X has value/desire Y, what will happen in the future?

## 6. Example Clip

We present our sample clip and questions based on the YouTube video clip from *The Devil Wears Prada* (The, 2014). Figure 2 presents eight screenshots from the video clip.


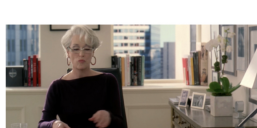






1.		Andy enters the office and Emily, Miranda's assistant greets her. Emily explains how important Miranda is in fashion and how hard competitive the job is. Emily assumes that Andy will not be fit for the job given her lack of experience with fashion.	5.		Miranda asks Andy what she is doing here. Andy replies that she thinks she would do a good job as Miranda's assistant.
2.		Emily and the others in the office scramble to prepare for Miranda's arrival in the office.	6.		Andy explains why she is interviewing for the position. Miranda is quiet and disrespectful.
3.		Emily introduces Andy to Miranda. Emily suggests to Miranda Andy is not able to complete the job.	7.		Miranda remarks that she doesn't like her fashion and ends the interview.
4.		Andy comes into Miranda's office to interview for a position.	8.		Another character enters the room and makes a negative comment about Andy.

Figure 2. Example Clip: *The Devil Wears Prada*. Description: A journalist named Andy goes to apply for a job as an assistant at a fashion magazine. Upon her arrival, Emily, another assistant, tells Andy she will not be hired due to her lack of fashion sense. Emily receives notice that Miranda, her boss, is coming, and the entire office scrambles to prepare for Miranda's arrival. When Miranda arrives, she interviews Andy. During their interview, Miranda is rude to Andy and suggests that Andy is not the right fit for the job.

### 6.1 Construct Map and Question Formulation

Here we provide example questions based on our example clip. For a full list of sample questions across each subdomain, see our appendix at <https://github.com/aivaslab/construct-map-appendix>.

#### 1. Perception Question

Subdomain: Facial Expressions, Difficulty: Hard

*How did Andy feel when Miranda told her "that's all"?*

- A. Upset and disrespected
- B. Happy and excited
- C. Uneasy and nervous
- D. Confused and curious

#### 2. Knowledge Question

Subdomain: Event Schema, Difficulty: Medium

*What does Andy expect from Miranda?*

- A. She expects Miranda to ask her questions about the position.
- B. She expects that Miranda will give her a hug.
- C. She expects that Miranda will be joyful to meet a new employee.
- D. She expects that Miranda will want to be her friend.



FRAMEWORK FOR A MULTI-DIMENSIONAL TEST OF THEORY OF MIND FOR HUMANS AND AI SYSTEMS

Perception				
Subdomain	Definition	Low Understanding	Medium Understanding	High Understanding
Prosody	Suprasegmental elements of language. Tone of voice, intonation, rhythm, stress, volume, etc.	Can understand exaggerated prosody.	Level 1 and can understand prosody use of volume and intonation at subtle levels.	Level 1 and 2 and picks up on subtle changes in prosody.
Body Language and Gestures	How a character presents their body as an indication of how they're feeling or what they're trying to communicate.	Can understand goal-directed gestures such as pointing.	Level 1 and can understand large body language signs such as shrugs, hunched shoulders, and arm movements.	Level 1 and 2 and can understand subtle changes in body language such as eye contact, body alignment, and nervous hands.
Facial Expression	The way a character makes their face look.	Can understand exaggerated expressions such as those in a cartoon (wide eyes, big smiles, big frowns)	Level 1 and can understand expressions expressing simple emotions such as happiness and sadness across different scenarios.	Level 1 and 2 and can understand complex, social emotions such as pride and shame.
Spoken Words	The words that a person speaks. Does not include how they speak it (aka prosody)	Uses some literal (less than 80%) language to make inferences.	Can track information from all concrete references.	Level 1 and 2 and can infer meaning from figurative language and allusions.
Non-Verbal Vocalizations	Laughter, grunts, gasps, groans, etc.	Can understand loud/exaggerated NVV such as loud signs and big laughter.	Level 1 and can sometimes (50 - 80% of the time) infer meaning from subtle NVV.	Can infer meaning from subtle NVV
Character Presentation	What a character looks like and is dressed like.	Can differentiate what characters are wearing/how they present themselves in character description	Can create labels for the character's presentation (for example "stylish", "like a cowboy", etc.)	Can use character presentation to infer traits of the character.
Setting	Identifying where a scene is taking place	Can differentiate between different settings that characters appear in.	Can create labels for the setting such as "office", "home", and "school".	Can use setting characters are in to infer about how characters should act.

Figure 3. Perception Construct Map. Definition column lists definition of corresponding perception subdomain (e.g., Prosody). Low Understanding, Medium Understanding, and High Understanding describe question difficulty for the corresponding subdomain.

Knowledge				
Subdomain	Definition	Low Understanding	Medium Understanding	High Understanding
Object Schema	Knowledge about how objects operate in the setting one is in.	Understands relevance of objects in common, everyday settings such as school and home.	Level 1 and can understand the role of objects in specialized settings.	Levels 1 and 2, and can understand combined schemas or schemas transformed from conventional presentation.
Event Schema	Knowledge about how events and scripts should precede given the event taking place.	Understand schema about basic everyday events, such as going to school or going to a restaurant.	Level 1 and can understand more specialized event schema such as weddings.	Level 1 and 2 and can understand combined schemas or schemas transformed from conventional presentation.
Behavior Schema	Knowledge about how people should behave given the setting.	Understands behavior schema for basic, everyday events such as going to school, being at work, or going to a party.	Level 1 and can understand behavior schema for specialized settings and events.	Level 1 and 2 and can understand combined schemas or schemas transformed from conventional presentation.
Personality Schema	Knowledge about how a specific person may act given their personality.	Understands personality schema for basic personality traits, such as kind or mean.	Level 1 and can understand personality schema for more complex personality traits, such as driven or arrogant.	Level 1 and 2 and can understand combined personality schemas or schemas transformed from conventional presentation.

Figure 4. Knowledge Construct Map. Definition column lists definition of corresponding perception subdomain (e.g., Event Schema). Low Understanding, Medium Understanding, and High Understanding describe question difficulty for the corresponding subdomain.

Reasoning (Abductive)				
Subdomain	Definition	Low Understanding	Medium Understanding	High Understanding
Emotions	Given that character X has emotion Y, why do they have that emotion?	Can understand directly stated reasons behind emotions from a character.	Can come up with a reason behind a character's emotions even when not directly stated.	Can come up with a reason behind a second-order emotion, even when not directly stated.
Intentions/Goals	Given that character X has intent/goal Y, why do they have that intention?	Can understand directly stated reasons behind goals and intentions from a character	Can come up with a reason behind a character's intentions and goals even when not directly stated.	Can come up with a second-order intention/goal, even when not directly stated.
Beliefs	Given that character X has belief Y, why do they have that belief?	Can understand the directly stated reasons behind a character's beliefs.	Can come up with a reason behind a character's beliefs even when not directly stated.	Can come up with a second-order belief, even when not directly stated.
Values/Desires	Given that character X has value/desire Y, why do they have that value/desire?	Can understand the directly stated reasons behind a character's values/desires.	Can come up with a reason behind a character's values/desires even when not directly stated.	Can come up with a second-order value/desire, even when not directly stated.
Reasoning (Prediction)				
Subdomain	Definition	Low Understanding	Medium Understanding	High Understanding
Emotions	Given that character X has emotion Y, what will happen in the future? ["what will happen" can include how they will feel in the future, what they will do in the future, what they will believe in the future, etc.]	Can make predictions based off of information directly mentioned.	Can make predictions that haven't been directly mentioned.	Can make predictions that haven't been directly mentioned about second-order emotions.
Intentions/Goals	Given that character X has intention/goal Y, what will happen in the future?	Can make predictions based off of information directly mentioned.	Level 1 and can make predictions that haven't been directly mentioned.	Can make predictions that haven't been directly mentioned about second-order intentions/goals.
Beliefs	Given that character X has belief Y, what will happen in the future?	Can make predictions based off of information directly mentioned.	Level 1 and can make predictions that haven't been directly mentioned.	Can make predictions that haven't been directly mentioned about second-order beliefs.
Values/Desires	Given that character X has value/desire Y, what will happen in the future?	Can make predictions based off of information directly mentioned.	Level 1 and can make predictions that haven't been directly mentioned.	Can make predictions that haven't been directly mentioned about second-order values/desires.

Figure 5. Reasoning: Abductive and Predictive Construct Maps. Definition column lists definition of corresponding perception subdomain (e.g., Emotions). Low Understanding, Medium Understanding, and High Understanding describe question difficulty for the corresponding subdomain.

### 3. Reasoning (Abductive) Question

Subdomain: Intentions/Goals, Difficulty: Easy

*Why does Andy want this job?*

**A. Andy needs more experience in journalism.**

B. Andy loves fashion.

C. Andy wants to meet Miranda because she is famous.

D. Andy needs to start earning more money.

## 7. Discussion

**Human Cognition and Learning.** The MD-ToM provides a useful framework for scene comprehension. The more detailed profile of the MD-ToM will allow for a thorough understanding of when aspects of ToM develop. For instance, we currently do not know of a ToM assessment that tests all the perceptual subdomains the MD-ToM has. Using the MD-ToM framework to assess neurotypical children and adults will reveal when fluency in each subdomain is achieved.

**Neurodivergent Populations.** The MD-ToM provides a framework for assessing neurodiverse populations and their ToM abilities. For instance, those with ASD are often found to have reduced ToM abilities (Kimhi, 2014). Our framework allows for a much more detailed analysis than current ToM tasks in assessing which aspects of ToM are reduced. This could provide important information to clinicians and families when considering interventions. For instance, if an adolescent with ASD takes MD-ToM assessment and it is revealed that their background knowledge and reasoning is at a medium and high level, but their perception is at a low level, especially in facial expressions and prosody, intervention time can be focused on increasing skills in the specific areas of need.

**AI System Design.** Our MD-ToM model has two-fold implications. Firstly, it will enable the development of ToM computational model from three dimensions: *reasoning*, *knowledge*, and *perception*. Current computational ToM models considers primarily False Belief tasks (see Eysenbach et al. (2016) for a list of works). Secondly, the MD-ToM framework may contribute to common-sense reasoning research in AI and NLP. Currently, researchers build AI systems by crowdsourcing data and developing models, usually along only one dimensions (reasoning or belief, see our related work at Section 4). Our framework will enable AI researchers to build cognitive AI models considering a holistic view of ToM across three dimensions.

## 8. Conclusion

In this paper we propose a theoretical framework for multi-dimensional ToM, MD-ToM. We propose three dimensions: knowledge, perception, and reasoning. Furthermore, we divide three dimensions into 19 subdomains and present construct maps defining high, medium, and low understanding of ToM across each subdomain. We also present example questions formulated using our MD-ToM construct map. Our proposed MD-ToM fills a research gap in cognitive science and AI, as real-world ToM reasoning rarely relies on just one type of perceptual judgment or making one particular type of inference. We hope our framework will be useful for cognitive scientist and AI researchers.

## 9. Acknowledgments

This work was funded in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A180171 to Vanderbilt University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## References

- (2014). The Devil Wears Prada. <https://www.youtube.com/watch?v=t4isatjZ0BM>.
- Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016). Neural module networks. *CVPR* (pp. 39–48).
- Baldwin, D. A., & Baird, J. A. (2001). Discerning intentions in dynamic human action. *Trends in cognitive sciences*, 5, 171–178.
- Baron-Cohen, S. (1997). How to build a baby that can read minds: Cognitive mechanisms in mindreading. *The maladapted mind: Classic readings in evolutionary psychopathology*, (pp. 207–239).
- Baron-Cohen, S. (2000). Theory of mind and autism: A fifteen year review. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience (2nd ed.)*, 3–20. New York, NY, US: Oxford University Press.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition*, 21, 37–46.
- Baron-Cohen, S., O’riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of autism and developmental disorders*, 29, 407–418.
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “reading the mind in the eyes” test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *J. Child Psychology & Psychiatry & Allied Disciplines*, 42, 241–251.
- Bauminger, N., & Kasari, C. (2000). Loneliness and friendship in high-functioning children with autism. *Child development*, 71, 447–456.
- Bauminger, N., Shulman, C., & Agam, G. (2003). Peer interaction and loneliness in high-functioning children with autism. *Journal of autism and developmental disorders*, 33, 489–507.
- Bora, E., & Berk, M. (2016). Theory of mind in major depressive disorder: A meta-analysis. *Journal of affective disorders*, 191, 49–55.
- Bosco, F. M., Colle, L., De Fazio, S., Bono, A., Ruberti, S., & Tirassa, M. (2009). Th. omas: An exploratory assessment of theory of mind in schizophrenic subjects. *Consciousness and cognition*, 18, 306–319.
- Bosco, F. M., Gabbatore, I., Tirassa, M., & Testa, S. (2016). Psychometric properties of the theory of mind assessment scale in a sample of adolescents and adults. *Frontiers in psychology*, 7, 566.

- Chen, Z., et al. (2020). Characterizing datasets for social visual question answering, and the new tinysocial dataset. *IEEE ICDL-EpiRob* (pp. 1–6). IEEE.
- Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160055.
- Curby, T. W., Brown, C. A., Bassett, H. H., & Denham, S. A. (2015). Associations between preschoolers' social-emotional competence and preliteracy skills. *Infant and Child Development*, *24*, 549–570.
- Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and speech*, *40*, 141–201.
- Damasio, A. R. (2004). Emotions and feelings. *Feelings and emotions: The Amsterdam symposium* (pp. 49–57). Cambridge University Press Cambridge.
- De Villiers, J. (2007). The interface of language and theory of mind. *Lingua*, *117*, 1858–1878.
- Duke, N., Pearson, P., Strachan, S., & Billman, A. (2011). Essential elements of fostering & teaching reading comprehension. *What research has to say about reading instruction*, *4*, 286–314.
- Eccles, J. S., Wigfield, A., et al. (2002). Motivational beliefs, values, and goals. *Annual review of psychology*, *53*, 109–132.
- Eysenbach, B., Vondrick, C., & Torralba, A. (2016). Who is mistaken? *arXiv preprint arXiv:1612.01175*.
- Fink, E., Begeer, S., Peterson, C., Slaughter, V., & de Rosnay, M. (2015). Friendlessness and theory of mind: A prospective longitudinal study. *British J. Developmental Psychology*, *33*, 1–17.
- Fletcher-Watson, S., McConnell, F., Manola, E., & McConachie, H. (2014). Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD). *The Cochrane Database of Systematic Reviews*, *3*, CD008785.
- Garner, P. W., & Waajid, B. (2008). The associations of emotion knowledge and teacher-child relationships to preschool children's school-related developmental competence. *Journal of Applied Developmental Psychology*, *29*, 89–100.
- Hammond, M. D., & Cimpian, A. (2017). Investigating the cognitive structure of stereotypes: Generic beliefs about groups predict social judgments better than statistical beliefs. *Journal of Experimental Psychology: General*, *146*, 607.
- Hamre, B. K., & Pianta, R. C. (2006). Student-teacher relationships.
- Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, *24*, 129–154.
- Hobson, P. (1993). Understanding persons; the role of affect. *Understanding other minds*, (pp. 204–227).
- Hutchins, T. L., Prelock, P. A., & Chace, W. (2008). Test-retest reliability of a theory of mind task battery for children with autism spectrum disorders. *Focus on autism and other developmental*

- disabilities*, 23, 195–206.
- Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *CVPR*.
- Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American journal of public health*, 105, 2283–2290.
- Kavanagh, D. J., Andrade, J., & May, J. (2005). Imaginary relish and exquisite torture: the elaborated intrusion theory of desire. *Psychological review*, 112, 446.
- Kimhi, Y. (2014). Theory of mind abilities and deficits in autism spectrum disorders. *Topics in Language Disorders*, 34, 329–343.
- Kintsch, W., & Walter Kintsch, C. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.
- Kirouac, G., & Doré, F. Y. (1983). Accuracy and latency of judgment of facial expressions of emotions. *Perceptual and motor skills*, 57, 683–686.
- Langlois, J. H. (1995). The origins and functions of appearance-based stereotypes: Theoretical and applied implications. In *Craniofacial anomalies*, 22–47. Springer.
- Leslie, A. M. (1987). Pretense and representation: The origins of " theory of mind." *Psychological review*, 94, 412.
- Li, J., Niu, L., & Zhang, L. (2022). From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. *CVPR* (pp. 21273–21282).
- Liddle, B., & Nettle, D. (2006). Higher-order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology*, 4, 231–244.
- Luengo, I., Navas, E., Hernández, I., & Sánchez, J. (2005). Automatic emotion recognition using prosodic parameters. *Ninth European conference on speech communication and technology*.
- Moore, C. (2013). *The development of commonsense psychology*. Psychology Press.
- Müller, E., Schuler, A., & Yates, G. (2008). Social challenges and supports from the perspective of individuals with asperger syndrome and other autism spectrum disabilities. *Autism*, 12, 173–190.
- Peterson, C., Slaughter, V., Moore, C., & Wellman, H. (2016). Peer social skills and theory of mind in children with autism, deafness, or typical development. *Developmental psychology*, 52, 46.
- Peterson, C. C., & Siegal, M. (2002). Mindreading and moral awareness in popular and rejected preschoolers. *British Journal of Developmental Psychology*, 20, 205–224.
- Phillips, W., Baron-Cohen, S., & Rutter, M. (1998). Understanding intention in normal development and in autism. *British Journal of Developmental Psychology*, 16, 337–348.
- Phutela, D. (2015). The importance of non-verbal communication. *IUP Journal of Soft Skills*, 9, 43.
- Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. *International conference on machine learning* (pp. 4218–4227). PMLR.

- Ramey, M. M., Henderson, J. M., & Yonelinas, A. P. (2022). Episodic memory processes modulate how schema knowledge is used in spatial memory decisions. *Cognition*, 225, 105111.
- Rashedi, R. N., et al. (2021). Opportunities and challenges in developing technology-based social skills interventions for adolescents with autism spectrum disorder: A qualitative analysis of parent perspectives. *Journal of Autism and Developmental Disorders*, (pp. 1–16).
- Ratcliffe, B., Wong, M., Dossetor, D., & Hayes, S. (2015). The association between social skills and mental health in school-aged children with autism spectrum disorder, with and without intellectual disability. *Journal of autism and developmental disorders*, 45, 2487–2496.
- Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., & Choi, Y. (2019a). Atomic: An atlas of machine commonsense for if-then reasoning. *AAAI*.
- Sap, M., Rashkin, H., Chen, D., Le Bras, R., & Choi, Y. (2019b). Social iqa: Commonsense reasoning about social interactions. *EMNLP-IJCNLP* (pp. 4463–4473).
- Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 107, 2408–2412.
- Sprong, M., Schothorst, P., Vos, E., Hox, J., & Van Engeland, H. (2007). Theory of mind in schizophrenia: meta-analysis. *The British journal of psychiatry*, 191, 5–13.
- Suhr, A., Lewis, M., Yeh, J., & Artzi, Y. (2017). A corpus of natural language for visual reasoning. *55th Annual Meeting of the Association for Computational Linguistics (2)* (pp. 217–223).
- Tager-Flusberg, H., & Sullivan, K. (2000). A componential view of theory of mind: evidence from williams syndrome. *Cognition*, 76, 59–90.
- Thoppilan, R., et al. (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.
- Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard university press.
- VandenBos, G. R. (2007). *Apa dictionary of psychology*. American Psychological Association.
- Wellman, H. M., & Lagattuta, K. H. (2004). Theory of mind for learning and teaching: The nature and role of explanation. *Cognitive development*, 19, 479–497.
- Wilson, M. (2004). *Constructing measures: An item response modeling approach: An item response modeling approach*. Routledge.
- Zadeh, A., Chan, M., Liang, P. P., Tong, E., & Morency, L.-P. (2019). Social-iq: A question answering benchmark for artificial social intelligence. *CVPR* (pp. 8807–8817).
- Zalta, E. N., Nodelman, U., Allen, C., & Perry, J. (1995). *Stanford encyclopedia of philosophy*.
- Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. *CVPR* (pp. 6720–6731).
- Zellers, R., et al. (2022). Merlot reserve: Neural script knowledge through vision and language and sound. *CVPR* (pp. 16375–16387).
- Zi, X., et al. (2020). Adapting educational technologies across learner populations: A usability study with adolescents on the autism spectrum. *CogSci*.