PSYCH 285: Final Paper

Robert Hawkins

December 13, 2016

Introduction

If there is a single unifying message to take away from the vast literature on theory of mind, it is that the mental representations we use to reason about other minds are broader, more flexible, and more multi-faceted than any single experimental task can hope to pull apart. Traditional false-belief tasks (e.g. Wimmer & Perner, 1983), for instance, have been immensely useful for probing the "layer" of our representation that allows another's beliefs to be held alongside conflicting information about the true state of affairs. They also serve as useful litmus tests for drawing comparisons between human and non-human primates (Krupenye, Kano, Hirata, Call, & Tomasello, 2016) or between younger and older children (Wellman, Cross, & Watson, 2001).

On the other hand, they are not particularly informative about other aspects of our representations, such as the conventions or norms we expect others to conform to (Lewis, 1969; Rakoczy, Warneken, & Tomasello, 2008) or the tradeoffs between costs and rewards we expect others to rationally navigate (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). We have also become more aware of other aspects that are *are* involved in task performance but were not originally intended to be. For instance, answering the critical question *"Where will Sally look for her ball?"* may depend on reasoning about the experimenter's intentions when they draw attention to object locations or take the communicative action of asking a pedagogical question (Strawson, 1964; Tomasello, Carpenter, Call, Behne, & Moll, 2005; Westra & Carruthers, 2017).

The Director-Matcher Task

Similar criticisms have recently been directed at the *director-matcher task* used to argue for failures of theory of mind use (Keysar, Barr, Balin, & Brauner, 2000;

Keysar, Lin, & Barr, 2003; Lin, Keysar, & Epley, 2010). While its history is shorter and less widely appreciated than the tradition of traditional false-belief tasks, it is the crux of one of the most influential arguments about adult theory of mind use to emerge over the past two decades: While adults are *capable* of using deploying theory of mind, they do not always use it reliably. In other words, adults are "reflexively mind-blind" and only engage in perspective taking through an effortful process.

To demonstrate this, Keysar and colleagues placed several objects in a grid between two players. One player, a confederate, gives instructions to the other player about which objects to move. Critically, some cells of the grid are occluded so that only the matcher can see the object. They found that when an instruction was ambiguous between a hidden object and an object in common ground, nearly three out of four matches reached for the hidden one at least once (Keysar et al., 2003). Since a participant accurately taking the director's perspective would realize that the director couldn't possibly be referring to an object they don't know about, this was taken as a *failure* of theory of mind. The same reference-game paradigm was used as a basis for a number of follow-up studies testing perspective-taking in children (Epley, Morewedge, & Keysar, 2004), among friends vs. strangers (Savitsky, Keysar, Epley, Carter, & Swanson, 2011), across Eastern and Western cultures (Wu & Keysar, 2007a), and in contexts of varying background information overlap (Wu & Keysar, 2007b).

While the basic effect seems robust, there has been substantial disagreement over whether it does, in fact, reflect a *failure*. Some of these counter-arguments point to methodological details. Hanna, Tanenhaus, and Trueswell (2003) argued that common ground is usually established incrementally – utterance by utterance – over a conversation rather than through an artificially-imposed environmental constraint and also that the original task lacked a necessary control condition where both the target and distractor object are in common ground. Another methodological counter-argument from Heller, Grodner, and Tanenhaus (2008) focused on the small set of items used and the wide variation in the relative fitness of the scripted word to the distractor versus the target (e.g. "tape" is more commonly used to refer to the hidden "roll of tape" than the shared "cassette tape"). Finally, Rubio-Fernández (2016) point out that participants who initially adopt a selective-attention strategy where they ignore occluded cells can perform perfectly on all measures while *not* using their theory of mind to consider the other's perspective at all. Together, these arguments make the case that the task is simply an unreliable test of theory of mind use, but do not make positive claims about how, when, or why theory of mind *is* used.

A separate line of criticism considers the *pragmatics* of the task in a way that reinterprets apparent failures on the task as successful uses of theory of mind (Rubio-Fernández, 2016; Hawkins & Goodman, 2016). One key question hinges on the use of confederates in language-based tasks (Lockridge & Brennan, 2002; Kuhlen & Brennan, 2013). A confederate's behavior often differs dramatically from a naive subject's behavior in the same context. The relevant properties of behavior in dialogue – latencies, prosody, eye gaze, lexical and syntactic alignment, choice of referring expression – are still largely unknown to experimenters and difficult to explicitly control. To the extent that their partner expects some properties of natural dialogue and hears something else, they may draw unexpected implicatures. Indeed, one conceptual replication of the task that used naive subjects in both roles found many more *overinformative* referring expressions than the confederate's script contained (Hawkins & Goodman, 2016). If the listener successfully used theory of mind reasoning to form expectations about how the speaker would behave in such as context, then the scripted utterances used by the confederate would naturally lead to errors. Using similar reasoning, Rubio-Fernández (2016) showed that listeners became suspicious of what was actually in common ground after hearing overinformative utterances, thus successfully using theory of mind reasoning.

A model of pragmatics in the director-matcher task

The most successful computational models synthesize the thicket of contradictory empirical evidence and informal theory-building into a set of precisely-stated hypotheses that can be compared on the basis of quantitative predictions. The current state-of-the-art in modeling theory of mind reasoning has not addressed the underlying pragmatics of the task (Baker, Saxe, & Tenenbaum, 2009; Kiley Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013), and the current state-of-the-art in modeling pragmatics in language (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Goodman & Frank, 2016) has not been applied to standard theory of mind tasks despite explicitly framing language understanding as social reasoning. In this section, I formalize the director-matcher task inside a rational speech act (RSA) model, where speakers produce informative utterances by reasoning about a rational listener agent who updates their beliefs about the world according to the literal semantics of the language. This allows us to test pragmatic accounts of task performance.

Previous theoretical advances in modeling pragmatics have typically hinged on 'lifting' inference over some aspect of communication that was previously assumed to be fixed. For instance, *manner implicatures* emerge from uncertainty over the true meanings of words in the lexicon (Bergen, Levy, & Goodman, 2016), nonliteral constructions like *hyperbole* and *irony* emerge from uncertainty over the topic of discussion (Kao, Wu, Bergen, & Goodman, 2014), and *prosody* and *ellipsis* emerge from uncertainty (in the form of a noisy-channel model) over the various ways that a speaker's utterance could have been corrupted (Bergen & Goodman, 2015). Note that these are not ad hoc extensions; sources of uncertainty that are not strictly necessary to explain a particular effect are often omitted for simplicity, but they all coexist within the same model.

Pragmatic critiques of the director-matcher task point to an additional source of uncertainty that has not previously been explored: uncertainty over a communication partner's visual context. Considering this uncertainty, of course, *is* theory of mind in action; our goal is to show how such a model built on social reasoning can produce the 'failures' found by Keysar et al. (2003) as well as the phenomena observed by both Hawkins and Goodman (2016) and Rubio-Fernández (2016). For simplicity, we consider the full 2 × 2 grid containing a single occluded cell, which was used by Rubio-Fernández (2016). We incorporate this source of uncertainty into both the speaker and listener models as follows:

1. A pragmatic speaker perceives an array of three objects $\mathcal{O}_S = \{o_i\}$, one of which is identified as a target t, but assumes that their partner sees an additional hidden object o_h that is not in common ground. The speaker has uncertainty over the identity of this object, represented by a prior distribution $P(o_h)$ which they marginalize over when deciding what utterance to produce:

$$S_1(u|t, \mathcal{O}_p) \propto \sum_{o_h} L_0(t|\mathcal{O}_S \cup \{o_h\})P(o_h)P(u)$$

2. A pragmatic listener, reasoning about this uncertain speaker, is presented with a full array of four objects $\mathcal{O}_L$, one of which is identified as hidden to the speaker, o_h. The listener, however, also has some uncertainty over whether the experimenter is deceptive, represented by a Bernoulli random variable d, such that they assign a small prior probability $P(d)$ to the possibility that o_h is actually in common ground. Upon hearing an utterance$u$, they update both their beliefs about which object is the target and whether the experimenter is being deceptive:

$$L_2(t, d|u, \mathcal{O}_L) = P(t)\left(S(u|t, \mathcal{O}_L)P(d)\right)^d \left(S(u|t, \mathcal{O}_L - o_h)(1 - P(d)\right)^{1-d}$$

Technical details are omitted here for brevity, but the full model is fully specified in WebPPL and available to run at `http://forestdb.org/models/keysar.html`.

We focus on two critical patterns of results produced by the pragmatic listener L_2. First, we consider how it responds to the scripted instruction “fish” when presented with a context containing two fish (a red one that is occluded and a blue one in common ground) as well as two other unrelated objects. While it is most likely to choose the red fish in common ground, it has a relatively high probability (43%) of choosing the hidden red fish, an effect Keysar et al. (2003) claimed was due to failure of theory of mind. This decision emerges in the model for two primary reasons. For one, a speaker with uncertainty over context could in principle be communicating about any object, and if they believed the hidden object was a red fish, the utterance ‘fish’ would have a high probability of being produced. To a lesser extent, context uncertainty makes the overinformative utterance ‘blue fish’ more likely than ‘fish’, hence the omission of the color modifier has the (subtle) implication of *not* the blue fish. This speaker phenomenon, while subtle under the parameter regime of the current implementation, produces the dynamic observed in the unscripted replication by Hawkins and Goodman (2016).

Second, we consider the listener’s posterior over the *deception* Bernoulli variable d. Just as Rubio-Fernández (2016) found that listeners were more likely to suspect deception after hearing an ‘overinformative’ utterance like ‘blue fish’ when only one fish was in common ground, so too does our pragmatic listener. The pragmatic speaker is relatively more likely to produce the additional color modifier if they see multiple fish than when they see a single fish and simply have uncertainty over the unknown object. Hence, our model simultaneously captures all three phenomena.

Discussion & Conclusion

Theory of mind use is intimately intertwined with language understanding; pragmatic reasoning is fundamentally social reasoning applied to communicative behavior. While traditional theory of mind tasks have always relied on language – a critique that was first raised by developmental researchers (Baillargeon, Scott, & He, 2010) – there is increasing recognition of the pragmatics inherent to these tasks (Westra & Carruthers, 2017). In this paper, we turned this critical lens on the popular director-matcher task and synthesized several recent results under a single model of social reasoning under uncertainty about shared context.

Just as no single theory of mind task is sufficient to pull apart the many aspects of our mental representations, no single level of analysis is sufficient to interpret the empirical evidence (Marr, 2010). We need mechanistic accounts that isolate distinct neural subsystems (Gallagher & Frith, 2003; Koster-Hale & Saxe, 2013); algorithmic accounts that focus on memory, attention, executive control, and other processing demands(Rubio-Fernández & Geurts, 2012; Apperly & Butterfill, 2009); and computational accounts that ask functional question: what problem is theory of mind useful for solving and what information must be included in our representation in order to solve it Baker et al. (2009); Tomasello (2009); Tomasello and Vaish (2013). Moving ahead, each of these lenses will be critical to understanding exactly what aspects of our interlocutor's mind we represent in pragmatic reasoning, why other primates seem unable to engage in such reasoning, and how it develops across the lifespan.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological review*, *116*(4), 953.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in cognitive sciences*, *14*(3), 110–118.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bergen, L., & Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. *Topics in cognitive science*, *7*(2), 336–350.

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*(20).

Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, *40*(6), 760–768.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in cognitive sciences*, *7*(2), 77–83.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818 - 829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43–61.

Hawkins, R. X. D., & Goodman, N. D. (2016). Conversational expectations account for apparent limits on theory of mind use. In A. Papafragou, D. Grodner,

D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th annual conference of the Cognitive Science Society.*

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–836.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25 - 41.

Kiley Hamlin, J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental science*, *16*(2), 209–226.

Koster-Hale, J., & Saxe, R. (2013). Functional neuroimaging of theory of mind. In *Understanding other minds: Perspectives from developmental social neuroscience* (3rd ed., pp. 132–163). Oxford University Press.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114.

Kuhlen, A. K., & Brennan, S. E. (2013). Language in dialogue: when confederates might be hazardous to your data. *Psychonomic bulletin & review*, *20*(1), 54–72.

Lewis, D. (1969). *Convention: A philosophical study.* Harvard University Press.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic bulletin & review*, *9*(3), 550–557.

Marr, D. (2010). *Vision : A computational investigation into the human representation and processing of visual information.* Cambridge, Mass.: MIT Press.

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: young children's awareness of the normative structure of games. *Developmental psychology*, *44*(3), 875.

Rubio-Fernández, P. (2016). The director task: A test of theory-of-mind use or selective attention? *Psychonomic Bulletin & Review*, 1–8.

Rubio-Fernández, P., & Geurts, B. (2012). How to pass the false-belief task before your fourth birthday. *Psychological Science*, *24*(1), 27-33.

Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology*, *47*(1), 269–273.

Strawson, P. F. (1964). Intention and convention in speech acts. *The philosophical review*, 439–460.

Tomasello, M. (2009). *Why we cooperate* (Vol. 206). MIT press Cambridge, MA.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(05), 675–691.

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual review of psychology*, *64*, 231–255.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 655–684.

Westra, E., & Carruthers, P. (2017). Pragmatic development explains the theory-of-mind scale. *Cognition*, *158*, 165–176.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103-28.

Wu, S., & Keysar, B. (2007a). The effect of culture on perspective taking. *Psychological science*, *18*(7), 600–606.

Wu, S., & Keysar, B. (2007b). The effect of information overlap on communication effectiveness. *Cognitive Science*, *31*(1), 169–181.