AMERICAN PSYCHOLOGICAL ASSOCIATION

American Psychologist

© 2024 American Psychological Association
ISSN: 0003-066X

2024, Vol. 79, No. 9, 1322–1336
https://doi.org/10.1037/amp0001440

The Detection of Automatic Behavior in Other People

Tomer D. Ullman and Ilona Bass
Department of Psychology, Harvard University

The standard model of theory of mind posits that we attribute mental states to other people to explain their behavior. However, what of cases in which we think the other person is being scripted, acting automatically with no goals or beliefs to recover? While a great deal of past work has distinguished between automatic and reflective behaviors in one's own decision making, here we argue that reasoning about automatic behavior in *other* people is an important and largely unexplored area in research into theory of mind. We report results from two studies ($N = 4{,}528$ total) that examine the detection of automatic behavior in others. In Study 1, we conducted a large-scale survey characterizing the ubiquity of rote interactions in people's daily lives. In Study 2, we showed participants short video clips from a variety of domains and found that people quickly and reliably attribute automaticity to others and that automaticity judgments are distinct from other related behavioral attributions. On the basis of our findings, we suggest that reasoning about scripted behavior in others is an important, frequent, intuitive inference and propose extensions to the current research in intuitive psychology to study it further.

Public Significance Statement
When interacting with another person, you may have the sense that the other person is acting in a scripted or automatic way. We find that such interactions are highly frequent and that the perception of automatic behavior is fast and consistent. We propose that research, which examines how we reason about the mental life of others, should be expanded to include reasoning about other people behaving automatically.

Keywords: automatic behavior, social cognition, theory of mind, cognitive flexibility, intuitive psychology

Supplemental materials: https://doi.org/10.1037/amp0001440.supp

Think of your thinking. When you are asked "what is 5 × 7?" the response is immediate, as if you consulted a lookup table in your head. However, when you are asked "what is 51 × 71?" the answer takes more time and is as if you are going through a step-by-step algorithm in your head. The distinction between habitual, lookup-tablelike behaviors and more planning-based, reflective reasoning applies to nearly all aspects of our lives: Navigating a new city to find a new café and striking up an

Editor's Note. Tomer D. Ullman received the 2024 APA Award for Distinguished Scientific Early Career Contributions to Psychology. In association with the award, Ullman was invited to submit a manuscript to *American Psychologist*, which was peer reviewed. The article is published as part of the journal's annual Awards Issue.

Philip David Zelazo served as action editor.
Tomer D. Ullman https://orcid.org/0000-0003-1722-2382
Ilona Bass https://orcid.org/0000-0002-8192-7646
All studies were preregistered, and the data are publicly available on the Open Science Framework at https://osf.io/6a7kw/?view_only=ad9a9f878ecb4d9bb762103e2b4a74fd.

The authors have no conflicts of interest to declare. Tomer D. Ullman is supported by the National Science Foundation, Science Technology Center (Grant CCF-1231216); the Defense Sciences Office, DARPA Machine Common Sense program; and the Jacobs Family Foundation. Ilona Bass is supported by the National Science Foundation, STEM Ed Postdoctoral Research Fellowship (Grant 2327447), and a Psychology Research Innovation Award from the Hodgson Fund at Harvard University.

The authors thank the members of the Computation, Cognition, and Development Lab for their helpful discussions. The authors especially thank Erik Brockbank and Paige Bost for their assistance with collecting the stimuli and building the experiments and Heather Kosakowski for discussions and comments regarding possible neural implementations of the detection of automaticity.

Tomer D. Ullman played a lead role in formal analysis, funding acquisition, investigation, methodology, project administration, resources, writing–original draft, and writing–review and editing and an equal role in conceptualization, data curation, and supervision. Ilona Bass played a supporting role in writing–original draft and writing–review and editing and an equal role in conceptualization, data curation, formal analysis, investigation, and supervision.

Correspondence concerning this article should be addressed to Tomer D. Ullman, Department of Psychology, Harvard University, 52 Oxford Street, Cambridge, MA 02138, United States. Email: tullman@fas.harvard.edu

impromptu deep conversation are examples of reflective behaviors. Taking the usual route to the usual coffee shop to order the usual coffee with the usual joke, these are scripts. The difference between automatic and reflective thinking in one's *own* decision making has been the topic of immense study. However, it is not our focus here.

Think of other people's thinking. If, while waiting for your coffee, you overhear a person snapping at their partner for being late, you may wonder what is going on in *their* head. Perhaps they are angry that this has happened too often before and feel disrespected. Perhaps they are upset about something else entirely and simply lashing out. Possibly a myriad of other inner lives come to your mind, which can reasonably explain the strained scene. How people attribute mental states to others has also been the topic of intense study, and it is also not our focus here.

Going back to the coffee shop one last time, consider the employee who took your order and asked "do you want milk with that?" Most likely, you do not reflect on their inner life at all. To the degree that you do, you do not reason about all the possible goals, beliefs, desires, and emotions that led them to ask you if you wanted milk. This is because there *are* no such mental states driving their action; there is only the script. The recognition that someone else is acting automatically, how we recognize this, and why it matters—that is our focus here.

Our aim is to call attention to this less studied aspect of intuitive reasoning about other people and to suggest through arguments and studies that it makes up a substantial portion of people's daily lives. We propose that reasoning about automaticity is as intuitive and basic as mental state inferences while being separate from the standard formal frameworks that are used to model theory of mind (ToM). Put differently, we suggest that reasoning about scripted behavior in other people is currently the "dark matter" of intuitive psychology—it is not well modeled or understood, but it makes up much of the topic of interest.

Our plan for the rest of the introduction is as follows: We first survey the separation between habitual and nonhabitual behavior in decision making. We then turn our attention to the standard framework of mental state attribution and ToM. We do this to highlight that reasoning about automatic behavior in others is related to, but separate from, these well-established lines of research. While this reasoning is less studied in comparison to the other lines, this is not to say it has not been the topic of study, and we consider recent relevant work. We then briefly detail our studies and main findings. With that roadmap in mind, we now turn to habitual and nonhabitual behavior in people's own reasoning and decision making.

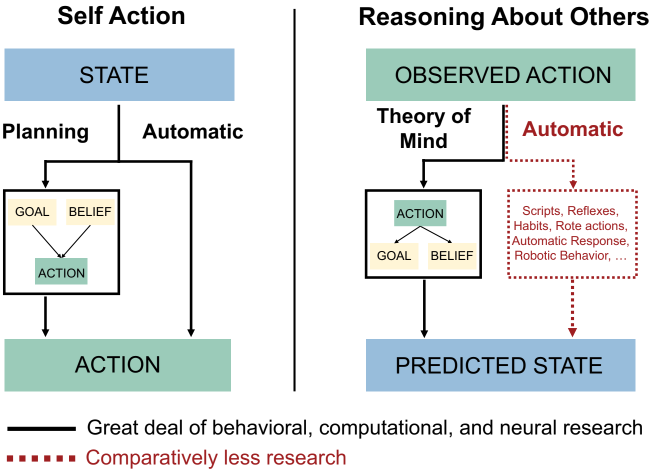
The broad separation between automatic behavior and more flexible reasoning reaches back to the origins of the fields of cognitive science and psychology (Thorndike, 1911; Tolman, 1948) and up to frameworks that remain influential in current times (Dickinson, 1985; Kahneman, 2011). This distinction has been extensively studied behaviorally and neurally in both humans and nonhuman animals (see, e.g., Balleine & O'Doherty, 2010; Killcross & Coutureau, 2003; Liljeholm et al., 2011; Tanaka et al., 2008; for reviews, see Botvinick, 2012; Dolan & Dayan, 2013). The neural and behavioral work in turn sets the foundation for influential computational models, which interpreted the distinction between automatic and flexible behavior in terms of model-free and model-based reinforcement learning (e.g., Dayan & Daw, 2008; Doya et al., 2002; Sutton & Barto, 1998). More recently, the reinforcement learning paradigm led to further neural, behavioral, and computational work that more finely distinguished how these different modes of behavior and thought trade off against one another and also challenged its basic dichotomy (Collins & Cockburn, 2020; Cushman & Morris, 2015; Gershman et al., 2014, 2015; Keramati et al., 2016; Lieder & Griffiths, 2020; Otto et al., 2013; Wunderlich et al., 2012). Such work has also underpinned current cutting-edge models in artificial intelligence (e.g., Hassabis et al., 2017).

It is hard to overstate the reach and influence of the behavioral, neural, comparative, cognitive, and computational work on habitual versus flexible reasoning in decision making. Much of this work deals with the "forward" direction, meaning it is focused on how people themselves act and think (see Figure 1, left). However, our interest here is in the "inverse" direction—how people reason about the actions of *others*. Here too, there has been a great deal of work across many fields, which we detail next.

We often explain and predict other people's observed behavior (actions, words, gestures) by referring to unseen mental states (beliefs, desires, intentions, emotions). The core principles of this intuitive psychology are cross-culturally shared, early developing, and present in nonhuman animals (see, e.g., Gao & Scholl, 2011; Gergely et al., 1995; Heider & Simmel, 1944; Kuhlmeier et al., 2003; Marticorena et al., 2011; Spelke, 2022; Spelke & Kinzler, 2007; Vallortigara, 2012), though it increases in sophistication throughout development (Tomasello, 2018; Wimmer & Perner, 1983). There are arguments about the nature, development, and assessment of this reasoning (see, e.g., Saxe, 2012), but a generally agreed-upon label for it is "ToM," and it is generally presumed that people reason about the actions of others by trying to figure out their mental states (see Figure 1, right). Hand-in-hand with empirical research, researchers have developed different computational frameworks that try to model and implement ToM. Such models can be broadly split into those that use an "inverse-planning" approach (e.g., Baker et al., 2009, 2017; Evans et al., 2016; Jara-Ettinger, 2019; Jara-Ettinger et al., 2016; Jern et al., 2017; Kryven et al., 2021; Shu et al., 2021; Ullman et al., 2009), those that rely on perceptual cues (e.g., De Freitas & Alvarez, 2018; Scholl & Gao, 2013; Scholl & Tremoulet, 2000), and more agnostic machine-learning approaches that try to learn the relevant

Figure 1

A Simplified Overview Situating Research on "Reasoning About Automatic Behavior in Other People" (Dashed, Red) in the Context of Existing Research (Solid, Black)


Self Action
Reasoning About Others
STATE
Planning
Automatic
GOAL
BELIEF
ACTION
ACTION
OBSERVED ACTION
Theory of Mind
Automatic
ACTION
GOAL
BELIEF
Scripts, Reflexes, Habits, Rote actions, Automatic Response, Robotic Behavior, …
PREDICTED STATE
Great deal of behavioral, computational, and neural research
Comparatively less research

Note. See the online article for the color version of this figure.

functions from data (whatever those are) or invert reinforcement learning (e.g., Hadfield-Menell et al., 2016; Rabinowitz et al., 2018).

Computational frameworks that try to capture theory-of-mind reasoning greatly differ in their commitments and mechanisms, but their usual starting assumption is that the thing to be modeled is how people attribute hidden mental states to others. Our departure point here is that in addition to this crucial inference, we need to account for the *separate* inference of reasoning about automaticity in others. We suggest that this inference is as fast and consistent as people's other intuitive social distinctions, such as whether an entity is animate or not (Rutherford & Kuhlmeier, 2013).

Certainly, the sense that other people are acting automatically has been explored in art, literature, and philosophy (e.g., Heidegger, 1967; Sartre, 1956). However, the topic of people's intuitive reasoning about automatic behavior in other people has not been the main focus of research in cognitive science to the same extent that automatic behavior has been in planning (the "forward" direction) nor to the extent that theory-of-mind reasoning has been in intuitive psychology. We strongly stress that we do not mean that there has been no research on this issue. Early work by Schank and Abelson (1977) studied how to build intelligent behavior in machines, placing "scripts" at the lowest rung. While their concern was primarily with the decision making of the agent itself (whether and how to deploy certain scripts), this work also drew on the observation that we often expect others to behave according to a script. More recently, Zawidzki (2013) argued that the use of "mind-reading" (in the sense of attributing propositional attitudes to others proper) was a much rarer operation than originally thought and that instead research should focus on *mind-shaping*, the manipulation of mental states in others.

In this research program, the intentional stance (Gergely et al., 1995) is taken to be much leaner and focused on more computationally sparse expectations about normatively correct actions. We ourselves accept the leaner interpretation of mentalizing in the sense that even infants can apparently use inverse planning to attribute goals and proto-beliefs to others, but we still separate this from the script approach that does not attribute mental states at all. In line with this stronger split, Gershman et al. (2016) directly noted the lack of reasoning about other people's habits in most computational approaches to ToM and used a combination of models and experiments to show that people can reason about other people's suboptimality by reference to habits. A few years later, Hawkins et al. (2021) extended the Rational Speech Acts model (Goodman & Frank, 2016) to account for a perspective-sharing task in which a speaker may be nonideal and specifically noted that "scripted" speaker statements used in previous tasks were perceived to be less informative than those solicited as "natural" in a follow-up task. Moreover, recently, Berke et al. (2023) expanded theory-of-mind reasoning to include inferences about the amount of mental effort another person puts into pursuing their goals, accounting for situations in which another person may be perceived as distracted or relying on memory to solve a puzzle.[1] We also note the recent focus on the development and acquisition of norms as a related topic (Hawkins et al., 2019; Schmidt & Rakoczy, 2023), in which research has examined and noted the fact that children assume

[1] We note that in their discussion, Berke et al. (2023) stated that their work "suggests an exciting space of social inferences about other people's thinking that has been previously neglected by classical ToM work focused on inferences about other people's beliefs and desires" (p. 6). Part of our point here is that we agree that this is both a lacuna and an exciting area for research.

some actions are taken and some artifacts are used the way they are because "this is how things are done" rather than strictly following from an effort-minimizing planning model. However, such norms could be seen as additional constraints in a planning, mental variable-based model rather than an alternative to a planning model all together ("John wants to use the flibbet, and the normative way to use the flibbet is counter-clockwise" still attributes a mental goal to John). On the topic of development, work has also examined children's understanding of habitual behavior in others as suboptimal (Goldwater et al., 2020). In a new and provocative preprint, Jara-Ettinger and Dunham (2024) proposed what they termed "The Institutional Stance," which they posit is on par with, interacts with, but is separate from the mentalistic stance. Several of the core tenants behind their argument for this stance apply to our argument here as well, including that the appeal to norms and scripts and the computations underlying this reasoning stand on their own rather than being inherited from either a general reasoning system or from a specific mentalizing system. However, while Jara-Ettinger and Dunham (2024) placed "institutions" at the heart of their ontology, our focus is on "scripts" and "automatic behavior" independent of that. To use a specific example, an institution such as a university or bank may dictate the kind of beliefs, desires, and other mental states a person should or does hold in a given situation, as a kind of ur-prior that establishes expectations about how a social situation is to unfold, but this would still be a case of constraining mental states, whereas our focus is on the lack of mental states to begin with. Certainly, an institution can also dictate the expected script in a given situation, but in that sense, it is separate from both scripts and mental states. Further, our own focus here is both on the importance of scripted and automatic behavior and on the detection, prevalence, and downstream consequences of noticing this behavior in others. We stress that all this is not to place our approach in opposition to the institutional stance, which we see as a separate, useful, and important line of research, motivated by many of the same considerations and signaling a move in cognitive science to mapping the terrain beyond ToM. In our own recent work (Bass et al., 2024), we examined the "why does this matter" angle of reasoning about other people's automatic behavior. We focused on pedagogy, as standard models of teaching do not capture the rote/reflective distinction, and the inference of automatic behavior on the part of the teacher that may have negative consequences for learning. We found that people naturally distinguished teachers acting in a more rote fashion and that such teachers were rated lower on a variety of pedagogical measures. We note though that while the topic of scripted behavior in pedagogy may have far-reaching consequences as pedagogy moves more and more toward automatic scaling, detecting scripted behavior and the resulting consequences extend far beyond pedagogy.

So far, we have argued broadly that reasoning about automatic behavior in other people should be a separate domain of research. This domain of research contains many questions, and it is beyond the scope of this article to go into all of them, just as one article cannot hope to cover all animacy detection or all ToM reasoning. Rather, our goal is to argue that it *is* a domain of research separate from forward planning and most current approaches to ToM (most of the introduction so far), to establish its frequency and robustness empirically (the next section), and to chart a course for further study (the focus of the General Discussion section).

In the following sections, we detail two studies in which we aimed to characterize the frequency and robustness of the detection of automatic behavior in other people. The first study was a large-scale survey among a representative U.S.-based population, showing that by people's own report, a significant amount of their daily interactions involves automatic behavior on the part of others. The second study used 90 short videos taken from a variety of domains, showing that people can quickly attribute perceived automaticity of others, that this attribution is consistent between people, and that it differs from other, related attributions (such as likability, interest, or engagement). After detailing our results, we discuss the main open directions in the study of the inference of automatic behavior.

Study 1: Survey on Everyday Automatic Behavior in Other People

As an initial exploration of reasoning about automatic behavior in others, we wanted to establish roughly how often people perceive other people to be acting automatically in a scripted way. To get at this, we designed a survey that asked people to consider how often they perceived others to be acting in an automatic, rote, or scripted fashion in interactions over the past week. We reasoned that interactions over a day may fluctuate more wildly, whereas interactions over longer periods of time are harder to recollect.

We emphasize upfront that any answers gleaned from such a survey represent a specific snapshot of a specific population at a specific time. Further, answers from such a survey do not correspond to the ground truth of how often a social partner was actually acting automatically (in a more "model-free" way). It is also likely that many mundane and rote interactions are forgettable, and so estimates given by people would be biased by selection, and a more ecological method would involve people being asked to report on interactions in real time as they happen.

This study and other studies reported here were approved by Harvard University's Institutional Review Board (IRB19-1861). All aspects of the studies—including sample sizes, analysis plans, and inclusion criteria—were preregistered prior to data collection (Study 1: https://aspredicted.org/4NG_BRD;

Study 2: https://aspredicted.org/Y2V_LH3). Study materials and de-identified data are publicly available on the Open Science Framework at https://osf.io/6a7kw/?view_only=ad9a9f878ecb4d9bb762103e2b4a74fd. Participants were recruited online (Peer et al., 2017) via the Prolific platform (https://www.prolific.com). Participants were restricted to those located in the United States, having completed at least 100 prior studies on Prolific (with an acceptance rate of at least 90%), and who did not take part in similar studies. All participants provided informed consent.

Participants and Methods

We recruited a total of $N = 3{,}000$ participants for this study. Thirty-eight participants were dropped from further analysis due to failing a comprehension check ($N = 28$) or an attention check ($N = 10$), leaving $N = 2{,}962$ for analysis. For these, we report the following demographics: $M_{\text{age}} = 41.6$ years; 55% reported as assigned female sex at birth, 44% assigned male, and <1% reported intersex or chose not to report sex; 72% reported their race as White, 10% Black or African American, 8% Asian, 1% Native Indian or Alaskan Native, 1% chose not to report race, and 8% reported other or chose not to report race; and for the highest education level attained, 12% reported high school diploma or equivalent, 20% some college but no degree, 40% bachelor or associate degree in 2- or 4-year college, 13% master's degree, 2% doctoral degree, 1% professional degree, 1% less than high school degree, and 1% chose not to report education level. These figures are broadly in line with representative U.S.-based survey demographics.

Participants were directed to an online survey hosted on Qualtrics (see Figure 2). After filling out a consent form and reading a brief overview of the task, participants were given the following definitions:

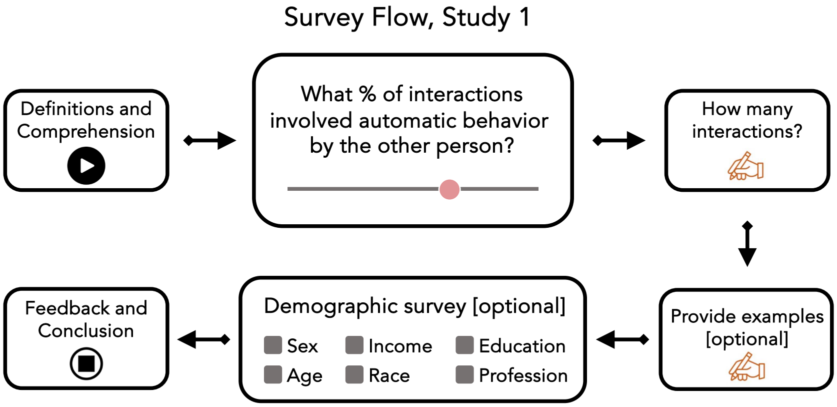
> When people act, they generally tend to use one of two ways of acting: 1. Automatic/Scripted: People are going through the motions, acting in a rote way, according to an internal script, using automatic behavior. 2. Reflective/On-the-fly: People are thinking through what they are doing or saying, flexibly responding to the environment.

Participants were then asked how well they understand the distinction, answering using a 5-point Likert scale ranging from *do not understand at all* to *understand completely*. Participants then completed an attention check, asking them to report which color word was placed on the right of an image. Having completed this warm-up, participants moved on to the main part of the study.

In the main part of the study, participants were asked to think back to the interactions they had with other people over the past week (this included all interactions with others, including in person, as well as also over the phone, via computer). Participants were then asked to give their best estimate for the percentage of interactions that included "automatic behavior" by the other person (meaning, the other person behaved in a way that seemed scripted). Participants reported their answers using a slider ranging from 0% (none of the interactions they thought about involved automatic behavior by the other person) to 100% (all of the interactions involved automatic behavior by the other person). In a separate screen, participants were then asked how *many* interactions they had with other people over the past week. Then, participants were invited to share an example of an interaction in which they felt the other person was behaving in an automatic fashion (this did not have to be from the past week, and the question was optional). Participants then filled out an optional demographic survey and were invited to share any feedback or comments they might have. Participants completed the survey in 4.5 min on average and were compensated 0.8 USD for their time.

Figure 2

Survey Flow for Study 1, Examining Automatic Behavior in Everyday Interactions


Survey Flow, Study 1
Definitions and Comprehension
What % of interactions involved automatic behavior by the other person?
How many interactions?
Provide examples [optional]
Demographic survey [optional]
Sex
Age
Income
Race
Education
Profession
Feedback and Conclusion

Note. See the online article for the color version of this figure.

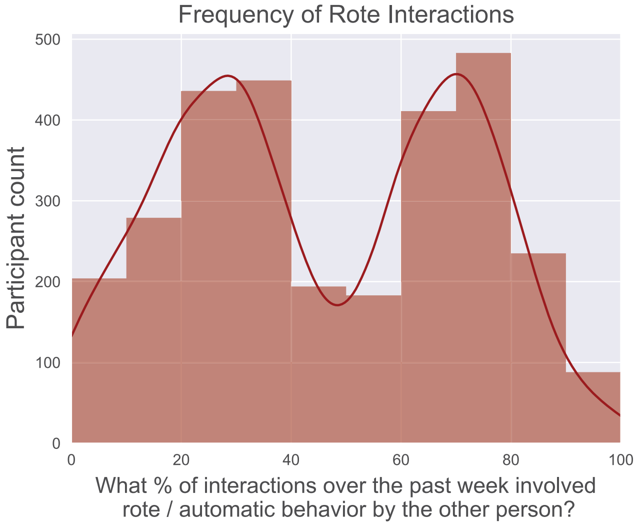
Results

On average, people reported that M = 44.5% of their interactions over the past week were such that the other person behaved automatically (95% CI [44.7%, 46.3%]). However, as shown in Figure 3, one should not take this average to be representative. The underlying distribution is strongly bimodal, with one mode around 30% and the other mode around 70%. Overall, a significant minority (45%) report that the majority of their interactions are such that other people behaved automatically or in a scripted fashion.

We again emphasize that while we surveyed a large and representative sample, these numbers represent the impressions of a specific slice of the U.S. population at a specific time and that they do not correspond to the ground truth in two ways. First, they refer to a remembered summary statistic that is likely biased by memorability. Second, they do not reflect whether the other person was *actually* acting in a scripted way (though this aspect of the ground truth is not of particular concern for our purposes here), though this is an interesting target for future research. All in all, we take the modes to reflect a sense of “somewhat more than half of my interactions are such” and “somewhat less than half of my interactions are such.”

As specified in our preregistration, we were not a priori committed to whether any of the demographic variables would be related to automaticity ratings. We found that there was no obvious predictor for which mode a participant would fall into. As detailed in the Supplemental Material, variables of sex, race, age, education, income, and profession do not meaningfully distinguish the two modes. Specifically for “profession,” we considered an exploratory analysis in which we examined whether a classifier trained on a high-dimensional embedding of the free-form responses given by participants could distinguish above- and below-average automaticity ratings. However, we found that it performed poorly (see the Supplemental Material for additional details).

Figure 3
Survey Results From Study 1



Note. N = 2,962. Participants reported on average that 44.5% of their interactions with others over the past week were such that the other people behaved in a way that seemed scripted or automatic. However, this average is a result of a bimodal distribution. See the online article for the color version of this figure.

The one factor that did show a small effect was “number of interactions,” such that a higher number of reported interactions corresponded to a lower frequency of perceived automatic behavior in others. This can be seen either by a Kolmogorov–Smirnov test applied to frequency-of-automatic-behavior distributions split by the median number of interactions (D = 0.07, p < .001) or by a linear regression using the log number of interactions as a predictor (r = −0.07, p < .001), and again see the Supplemental Material for details.

Beyond ratings, participants also provided free-form examples in response to an optional prompt asking for examples of interactions in which they felt the other person was behaving in an automatic fashion (not necessarily from the past week). After cleaning up responses such as “no” and “can’t think of any right now,” we were left with approximately 1,500 responses, which we have made publicly available at https://osf.io/skgpz?view_only=ad9a9f878ecb4d9bb762103e2b4a74fd. Unsurprisingly, many of the responses refer to interactions with customer service (at a shop or grocery store, on a cruise, with a cashier, at the gym, at a coffee shop, over the phone, fixing utility bills, telemarketers). These accounted for at least 35% of the responses, including examples such as:

> A checkout clerk at a local grocery store always seems to react in an automatic manner, despite knowing her for some time.

or

> At the convenience store, the owner who I see a lot is friendly, and follows the same social script every time I see him. He always says, “Hi buddy,” when I come in, asks if I want anything else and found everything I need, tells me the total and asks, “receipt?” after I pay. It seems automatic as he never really deviates from this script and I see the same with other customers.

People also mentioned office situations, dealing with coworkers and managers, medical situations, and interviews, as follows:

> My boss acts this way. He says the same thing every day without going out of the norm and its almost like he is a machine.

> Lots of interviewees prep in advance, so sometimes I feel like the answers I’m given are scripted or pre-thought out.

> The people that I have mostly encountered have been at the doctors office. I feel that in the doctors office, the behaviors are all scripted to

> make you believe that they care about your well being. To me it seems fake and scripted.

Going beyond mundane office, business, medical, and service situations, many people reported examples that were more intimate in nature. People reported automatic behavior on the part of friends, psychiatrists, parents, children, partners, and more. For example:

> Anytime I am talking to my wife, but she is not looking at me and she is looking at her phone feel automatic. She is responding, but I don't think she is comprehending what I am saying she is just agreeing with me to get me to stop talking so she can focus on her phone.

> My husband regularly responds to me with his "I'm paying attention script" while he's actually paying attention to his phone. A series of ahs, mmhmms, and oh?s.

Taken together, the survey results suggest that the perception of automatic behavior in others is a pervasive part of daily social life. Nearly half of the participants reported that a majority of their interactions were such, and the other half reported a significant amount of their interactions (about a third) were such. However, this survey relied on people's approximate recollections, and it is possible that people understood our explanations of automatic/scripted behaviors differently. In the next study, we turn to a more direct examination of the perception of automatic behavior in others by showing people the same set of stimuli and examining the consistency of their responses when asked about automatic behavior in others.

Study 2: Perception of Automaticity in Video Stimuli

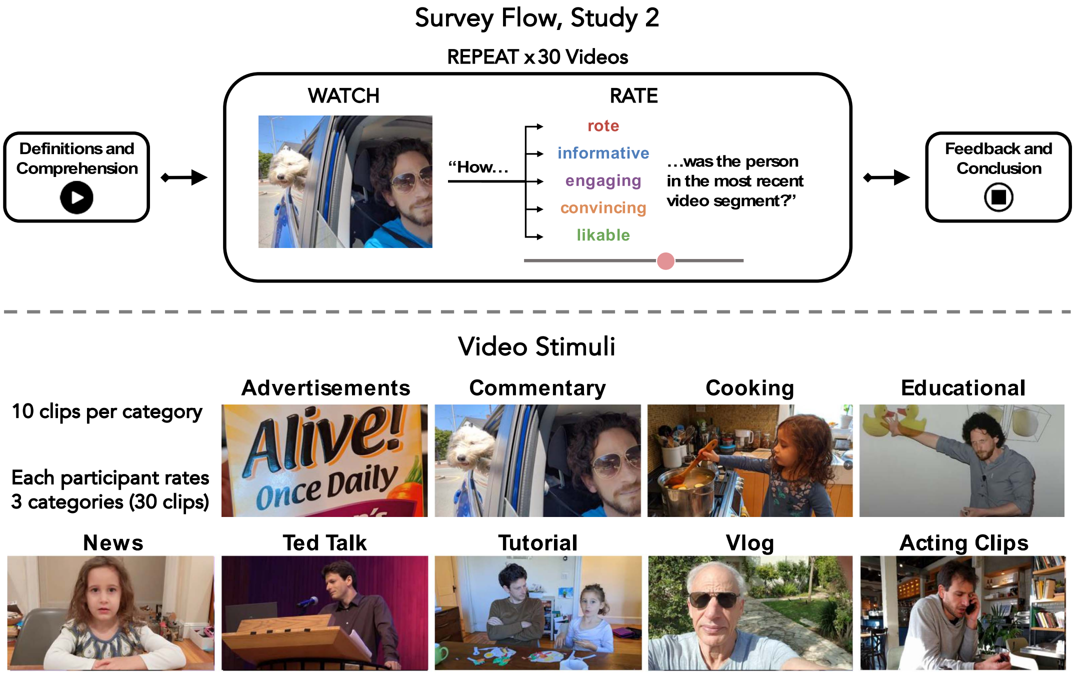
Our next study examined the robustness and reliability of people's perceptions of automaticity in others and how these judgments may be distinct from other aspects of behavior. We compiled a large set of videos, broadly representative of the kinds of content people might engage with in their day-to-day lives, and asked participants to evaluate how (a) rote, (b) informative, (c) engaging, (d) convincing, or (e) likable the speaker in the video was. We were interested in the nonrote judgments as potentially related to, but separate from, the detection of automaticity. See Figure 4 for an overview of the method used in Study 2.

Participants and Materials

Participants were U.S.-based adults recruited through the Prolific platform (https://www.prolific.com). We recruited a total of N = 1,566 participants for this study. Fifty-nine additional participants were dropped and replaced due to failing an initial comprehension check (N = 10) or inattentive

Figure 4

Study 2: Survey Flow (Top) and Domains Used in Stimuli (Bottom)



Note. Top: Survey flow for Study 2, examining the perception of rote/automatic behavior in short video stimuli, taken from a variety of sources. Bottom: The different domains used for the video stimuli, there were 90 videos in all (10 videos × 9 domains). Note that the subfigures in this image are for illustrative purposes only and are not from the videos originally used in the study. These images are published with permission. Links to the original YouTube clips can be found in the Open Science Framework additional online materials (https://osf.io/6a7kw/?view_only=ad9a9f878ecb4d9bb762103e2b4a74fd). See the online article for the color version of this figure.

patterns of responding over the course of the study ($N = 49$). For the 1,566 participants included in the analysis, we report the following demographics: $M_{\text{age}} = 40.3$ years; 54% reported as assigned female sex at birth, 45% assigned male, and 1% chose not to report sex; and 67% reported their race as White, 15% Black, 7% mixed race, 6% Asian, 3% other, and 2% chose not to report race. The median completion time for the study was 16.25 min, and participants were compensated 3.35 USD for their time.

We compiled a set of 90 videos from YouTube (https://www.youtube.com). There were 10 videos in each of nine categories: advertisements, commentary videos, cooking videos, educational videos, news clips, Ted Talks, tutorials, vlogs (“video blogs”), and acting clips. Each video was between 8 and 48 s long; videos longer than 20 s were broken up into approximately 10-s clips. Video clips were selected by searching for keywords from the category names (e.g., interviews, news) and browsing the “Trending” tab across multiple days. The researchers compiling the clips also perused their own and colleagues’ watch histories to ensure a broader sample.

Method

After filling out the consent form and reading a brief overview of the task, participants were randomly assigned to make one of the five judgments listed above (automatic $N = 309$; informative $N = 320$; engaging $N = 316$; convincing $N = 309$; likable $N = 312$). Participants were then given a working definition of the judgment they would be making (see the Supplemental Material for the full text of the definitions provided for each judgment). Participants were asked to indicate whether they understood the provided definition on a 5-point Likert scale, from 1 (*definitely not*) to 5 (*definitely*). Data from participants who answered three or lower on this comprehension check were dropped and replaced.

Participants then moved to the main part of the study, watching video clips one at a time. Each participant was randomly assigned to view all 10 videos within three random categories. For example, one participant might see vlogs, educational videos, and Ted Talks, while another might see educational videos, acting clips, and advertisements. The order in which videos were presented was fully randomized. Depending on their assigned condition, participants rated each clip based on how automatic, informative, engaging, convincing, or likable the person in the video was on a scale from 0 (*not at all*) to 100 (*extremely*). We also measured participants’ response time (i.e., the amount of time between the end of each video clip and when the judgment was submitted).

Transparency and Openness

All studies reported here and in the following sections were preregistered, including sample sizes, methods, exclusions, and analyses. The preregistration details as well as the data and research materials are publicly available on the Open Science Framework at https://osf.io/6a7kw/?view_only=ad9a9f878ecb4d9bb762103e2b4a74fd.

Results

We first examined overall automaticity impressions by the different video categories. Participant ratings for the domains roughly distinguished three groups with high variance within domains. These groups were “Advertising” at the high end, “Vlog” and “Commentary” at the low end, and every other group in the middle (see Supplemental Figures S4 and S5 for additional details about this analysis). The higher overall ratings of “Advertising” versus the lower overall ratings of “Vlog” are hardly surprising and could indicate a combination of top-down expectations and specific cues. Of more import is that the domains overall broadly overlap and are a useful validation check that the results below are not driven solely by domain.

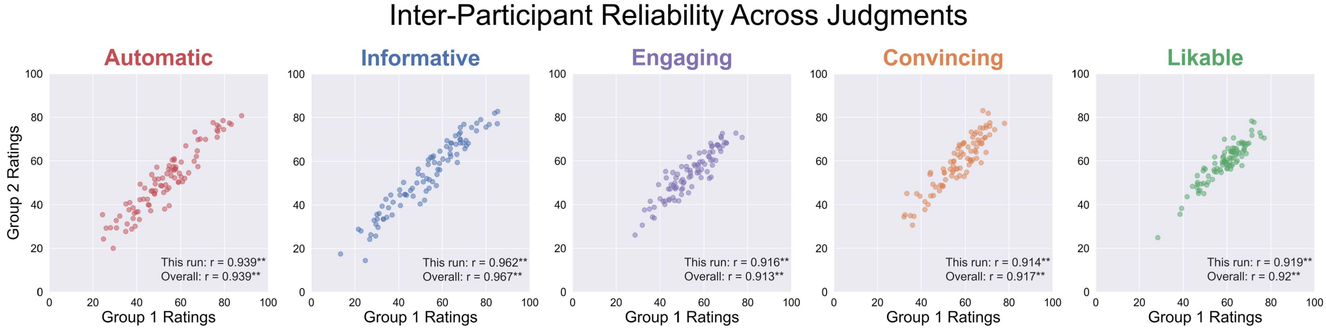
Next, we investigated how consistent perceptions of automaticity were across participants: Did people tend to agree about how scripted the people in the videos seem? We examined the interparticipant agreement by performing 1,000 runs of split-half correlations on the average rating for each video across participants. The average correlation between randomized participant half-splits was $r = 0.939$, 95% CI [0.938, 0.940], indicating a high level of agreement between participants overall. Participants were similarly reliable in their judgments of automaticity as they were in their evaluations of the speakers’ other attributes (see Figure 5). This reliability was not driven solely by domain, and the interparticipant reliability was high and significant within each domain as well (see Supplemental Figure S6 for automaticity ratings broken down by video category).

Judgments of automaticity were also just as fast as the other four behavioral attributions we asked participants to make. On average, it took participants 5.63 s after the end of each video clip to submit their judgment about how rote the people in the video were. As shown in Figure 6, this was similar to the amount of time it took participants to rate the speaker’s informativeness ($M = 5.69$ s, $t = 0.36$, $p = .720$), engagingness ($M = 5.70$ s, $t = 0.35$, $p = .730$), convincingness ($M = 6.27$ s, $t = 2.32$, $p = .020$), or likability ($M = 6.03$ s, $t = 2.15$, $p = .031$). For a more detailed figure of reaction time as a function of rating, see Supplemental Figure S7.

While automaticity ratings were as consistent and fast as other kinds of related attributions, we found that automaticity ratings did not correlate with the other ratings we examined, with the exception of informativeness (see Figure 7). This suggests that “automaticity” can be distinguished from other attributions and that a person may be judged as quite engaging (or convincing, or likable) and yet appear robotic and scripted. Regarding informativeness, we found that this attribution

Figure 5

Interparticipant Reliability/Agreement Across Judgments in Study 2



Note. Each dot represents one of 90 videos. Subfigures show the result of a single split-half correlation, breaking down participants into two groups and examining whether the average ratings of one group correlate with the other group. Text on the subfigures shows the correlation for this run, as well as the average of 1,000 such runs. See the online article for the color version of this figure.

was positively related to automaticity ($r = 0.446$), accounting for about 20% of the variance. This relationship may have also been driven by the particular categories of videos that we selected for use in this study: While it held qualitatively for all but one of the domains, it held quantitatively only for news and commentary (see the Supplemental Material and Supplemental Figure S8 for more details about the breakdown of automatic vs. informative correlations video category).

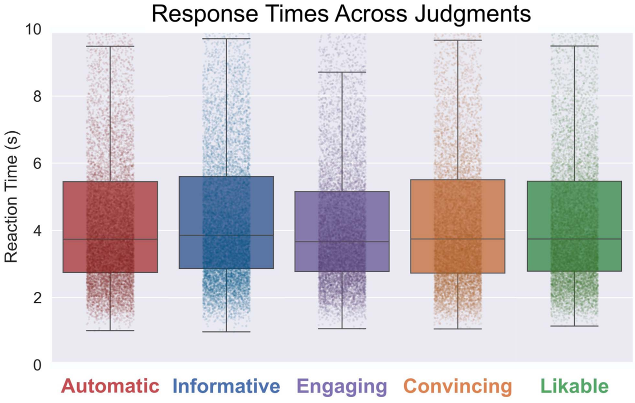
Together, these results demonstrate that people quickly attribute automaticity to others' behavior, that this attribution is reliable between individuals, and that automaticity judgments are distinct from other related behavioral attributions.

General Discussion

The salesman's pitch, the politician's speech, the teacher's drone, the nurse's concern, the beloved's "oh?" We are daily spectators to actors, going through their lines. While much research across disciplines has examined the split between scripted and improvised behavior in decision making, the research into ToM often takes as its starting point the notion that people see others as being driven by mental states such as goals and beliefs and focuses on the specifics of how such goals and beliefs are estimated. If, however, other people are acting automatically, there simply *are* no goals and beliefs to recover. There is only the script.

Figure 6

Participant Reaction Times in Study 2, Showing the Length of Time It Took Participants to Provide Different Evaluations of Speakers in a Short Video Clip, by Type of Evaluation



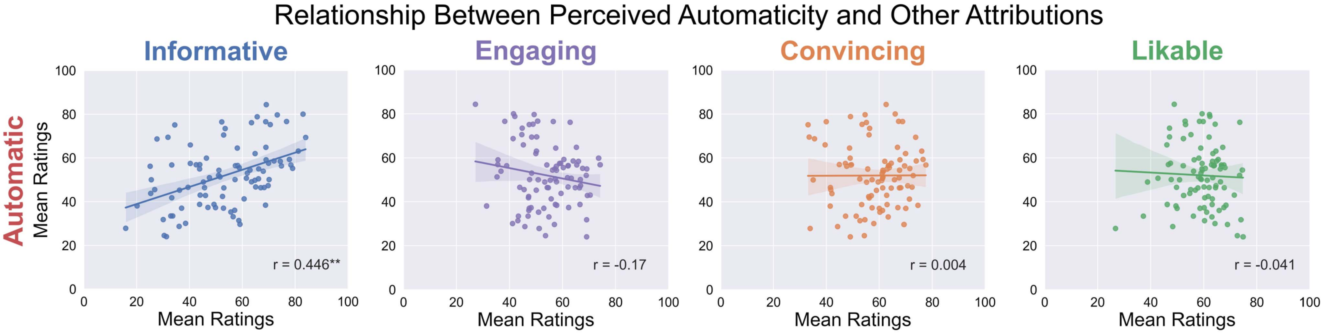
Note. Each dot represents the reaction time of a single rating. See the online article for the color version of this figure.

We argued that reasoning about scripted behavior in others is an important but mostly uncharted territory in cognitive research. Our empirical findings support the notion that this territory is also quite large, with nearly half of the respondents in a large representative survey reporting that a majority of their weekly interactions involved automatic behavior on the part of others. People gave many examples of such behavior in mundane settings including office settings, customer service, grocery stores, phone conversations, medical appointments, and so on. However, people also reported such behavior in more intimate relationships, including friends and loved ones. Beyond self-reported summaries and examples, we also examined people's attributions of automatic, rote behavior to other people more directly, using short video clips taken from a variety of domains. We found that people made such attributions as quickly and reliably as, but separate from, other related attributions. Together with the survey, the results paint a picture of a simple, intuitive, daily kind of behavior that is outside the scope of standard belief–desire inference.

While we examined and characterized the inference that other people are behaving in a scripted fashion, we did not touch on the "how." The speed and consistency with which people reasoned about automaticity in Study 2 suggests that this process is relatively automatic itself, but this should not be taken to suggest that it is strictly a bottom-up process. As touched on in the introduction, a useful analogy here is that of animacy detection (Gao et al., 2019; Scholl & Tremoulet, 2000; Spelke, 2022; Tremoulet & Feldman, 2006). Even at an

Figure 7

Correlations Between Ratings of Automaticity (y-Axes) and Other Evaluations of Speakers (x-Axes) in Short Video Clips in Study 2



Note. Only "informativeness" showed a significant relationship. See the online article for the color version of this figure.

early age, humans can distinguish animate and inanimate beings, and they do so on the basis of a collection of both bottom-up features ("starts from rest," "has eyes"), as well as more general and top-down principles ("violates physical laws"). Such an inference is necessary to shunt additional processing (e.g., if something is an agent, it may be moving toward a goal; if it is an object, it is morally permissible to smack it, and so on). In other work (Bass et al., 2024), we established repetition, attention, and the lack of verbal disfluencies as three possible cues for such an inference. However, it is likely that people distinguish others as acting in a scripted way through a host of bottom-up features including tone, eye contact, pitch modulation, the lack of verbal disfluencies, timing, and so on. Such prosody, timing, and fluency features have been studied extensively in children and adults as variables that can be used to infer mental states including processing difficulty, deception, intention, relative value, social connectedness, and knowledge (Arnold et al., 2007; Fox Tree, 2002; Heller et al., 2015; Kidd et al., 2011; Loy et al., 2017; Orena & White, 2015; Richardson & Keil, 2022; Templeton et al., 2022; Zhang et al., 2023) but less so in the study of the lack of mental states altogether. A possible point of departure then would be to take clips rated for automaticity and their associated ratings in Study 2 and train classifiers on acoustic features (cf. De Pinto et al., 2020) or multimodal features (cf. Williams et al., 2018). However, we note that any perceptual cue found this way would be "defeatable" (just as animacy cues are), and people also likely use top-down reasoning about a situation to judge when someone is being scripted. By "defeatable," we mean that such perceptual cues are neither necessary nor sufficient, just as "has eyes" as a feature for animacy does not account for dolls (inanimate) or eyeless creatures. Similarly, a well-trained actor may practice their "uhhs" and "umms" to give a sense of not being scripted even though we know they are, while others may intentionally and reflectively engage in behavior that appears outwardly scripted and robotic.

Relevant to the point above about fast, bottom-up detection, we caution that given our current methods, our results regarding the speed of judgments of automaticity relative to other judgments need to be further established. Specifically, it is possible that people are coming to an inference over an attribute (e.g., engagement) faster than automaticity and have already made up their mind about it as the clip is unfolding but that the need to watch the clip all the way through masks this difference. While our upper bound is still quite short and in line with the claim that automaticity judgments are relatively rapid, a different test of the timing of automaticity judgments would be to allow people to give an attribute judgment at any point while watching a clip, which could show potential differences in relative timing between attributions (and would also give a better estimate of the time to accumulate evidence). Another option would be to show snippets of different lengths from the same clip to different people and estimate how much of the variance in answers is determined by the first snippet versus the first two snippets and so on up until the full clip.

Another target for future work would be the downstream consequences of detecting that someone else is behaving in a scripted fashion. This is a good time to note that the realization that someone else is acting automatically does not necessarily lead to a negative evaluation. Just as it makes sense from a resource-rational perspective to cache many behaviors, and it would not do to calculate "7 × 5" from scratch every time, it is to be expected that there are many situations in which scripts are called for. In such situations, it is expected that other people would indeed behave in a scripted way, and deviations from it would be seen as odd. If a cashier asks you "how are you today?" and you answer "oh, thank you so much for asking, I'm not so great. Actually, my cat has been sick all night, and … ," the cashier would be justified to think something is wrong with you. They did not ask you how you were, they said their lines, and you were supposed to say yours. The potentially more consequential and frequent way that scripted and nonscripted behaviors get crossed is for a person who is ostensibly supposed to be nonscripted to suddenly seem scripted. Qualitatively, many of the participants expressing frustration or disdain for the detection of automaticity in our survey seemed to be frustrated about this sort of thing. A caretaker or loved one simply going through the motions without being driven by a goal or desire to interact is jarring, even if they are saying the right things. A teacher may be giving relevant, informative, useful information (Bonawitz & Shafto, 2016), but if they are seen as having pressed "play" on an internal tape recorder, listeners may stop paying attention (Bass et al., 2024). A politician may address their crowd in energetic tones, pausing in just the right place to tell just the right anecdote in just the right way, but if the whole thing seems *too* right, they may lose out to an opponent that is distracted, meandering, and baffling but is at least not a *robot*.

We mentioned several times that the detection of automaticity is outside standard computational models of ToM, and it is useful to consider how such models should be expanded to account for it. We can imagine at least three possibilities for such an expansion, which also set the stage for different neural proposals. On the first possibility, the initial inference people make in interactions is an animacy judgment, deciding whether the entity they are interacting with a person or an object. Anything passing for a person would then be handled by a "ToM" module, *including* reasoning about other agents that behave in scripted ways. Such an expansion would require us to reevaluate what ToM even means, if it can include reasoning about people acting not according to mental states such as goals, beliefs, and desires. However, it may be that past the realization that another agent is a person, there is the realization that they are scripted, which requires other processing than ToM. On the second possibility, it is possible that reasoning about scripts in others is parasitic on our own scripted behavior in self-action, similar to the proposal that mental state attribution relies on running a "simulation" of our own decision making (see, e.g., Gallese et al., 2004; Nichols & Stich, 2003; Rizzolatti et al., 2001). As a final possibility, it may be that following the realization that someone else is scripted, such behavior is handled neither by a parasitic simulation of our scripts nor by standard ToM but by a different module entirely. Such reasoning may help explain why we naturally consider other people who are scripted as "robotic," "not really there," "puppets," "nonplayable characters," and so on and would require a separate "theory of machines" (cf. McCoy & Ullman, 2018). We note that the early shunting of processing to different modules has its analogs in other domains, like the abovementioned early-developing ability to distinguish between objects and agents (Spelke, 2022; Spelke & Kinzler, 2007) and how people apply this classification before using more domain-specific reasoning. The idea is then that people first classify whether a person is acting in a scripted/rote/automatic fashion (as opposed to a more reflective fashion). If another person is scripted, further considerations of their goals and beliefs need not apply, and full-blown ToM computations need not get off the ground (in line with proposals such as Zawidzki, 2013). Much like the inference of animacy, the inference of automaticity likely can happen either in a bottom-up fashion (using perceptual cues, e.g., a person's tone and affect) or in a top-down fashion (using general world knowledge, e.g., that interactions with a waitress are scripted or that actors in the theater have memorized their lines or being informed by a confidant that a suave presenter has given the exact same talk elsewhere).

Our current data do not adjudicate between the possibilities spelled out above about the right way to expand the current models of ToM to encompass reasoning about automaticity. However, we do take the current stance that the third option is the most likely, in which ToM and reasoning about automaticity reign over separate domains that nevertheless have traffic going between them. To this view, one might object that in our data and our motivating examples, one cannot separate the scripted or automatic behavior of people from *some* mental states. For example, it is possible that the barista that asks for your milk preference is acting according to a script but nevertheless is being driven by mental states such as the *intention* to do their job well or their *desire* to take on a role in a script.[2] To this, we would agree that the current data in principle do not rule out such an option but that it is nevertheless the one that carries the explanatory debt. Such an account posits additional, seemingly unnecessary mental variables, so it would fall short by the standards of an economical view of cognition. Similar considerations could apply to theory-of-mind models as well, in which every desire or goal could in principle also carry the second-order intention of having their goal or the belief that one has that goal. While such second-order intentions may have their role in some situations, proving their existence or usefulness has been no mean feat (as a useful positive example, see the work of Kleiman-Weiner, 2018). Beyond this, we should again be careful about distinguishing between the forward-planning direction (in which it may or may not be the best explanation for decision-making researchers to say that the decision process in the barista happened via an initial mental variable representing intention, from which the script unfolded) and the inverse-planning direction (in which it seems a priori unlikely that Yuki the customer thought about the barista's intention at all when reasoning about the barista's behavior).

While we take the stance that a separation of modules will turn out to be the more likely description of the relationship between ToM and reasoning about automaticity, that does not mean they cannot interact in some cases. Under the relevant circumstances, one can understand others as having the *belief* that a script would be useful, or the *desire* to not act in a

[2] We thank an anonymous reviewer for this point.

scripted way, or a *goal* of deploying a specific script. Such interactions would be not unique to ToM and automaticity but are expected between many mental modules. A person can understand the goals and desires of others as constrained by physical realities, normative considerations, social pressures, and so on. However, such interactions are not strictly necessary nor does their existence suggest that physical reasoning, normative expectations, social consideration, scripted reasoning, and intuitive psychology are all one thing. Such modules are usefully studied as separate from, equal to, and interacting with each other.

In claiming that automaticity was a separate domain from ToM, one could continue the move even further and argue that this domain itself is not monolithic and holds divisions within subdivisions. While "model-based" and "model-free" have been useful dichotomies of action in planning, perhaps the labels "habitual," "rote," "automatic," "scripted," or others should not apply interchangeably to us reasoning about the behavior of others when they put on a seat belt (habitual, thoughtless action that is not a social "script") compared with the behavior of a clerk who wishes us a nice day (a move in a social script that does not express genuine concern), with the smooth performance of a candidate delivering a lecture (a specific routine developed through great effort to come off as natural but being too smooth), and with yet further cases (and we thank an anonymous reviewer for this point). To this, we would agree that our use of "automaticity" here has at times been nebulous and that there may be subpartitions within it for reasoning, for example, about other people's habits compared with social scripts. However, we would suggest that many of the seeming complications and subflavors of automaticity do not themselves reflect different stand-alone domains but rather the interaction of the detection of automaticity and reasoning about scripts with other processes.

One may wonder, if reasoning about scripted or automatic behavior is separate and equal to ToM, why people seem so ready to attribute mental states in cases where they do not apply. For example, Mazar and Wood (2022) showed that people seem to underattribute behavior to habits, and Cushman (2020) nicely summarized much of the "rationalization" work as being people attributing goals and beliefs where none apply and then bringing their own goals and beliefs in line with those and for good reason. To this, we would broadly answer that the exact separation of automaticity and ToM will take much more research to delineate. We would more specifically answer that much of that work was concerned with people's explanation of their own behavior and that the pragmatics of the questions ("Why did you do that?") may suggest to people that their response should be in a belief–desire schema, even if they have no access to the underlying actual reasons. It is an interesting possibility that if offered the explicit option of explaining behavior in terms of scripts and automatic behavior people would make ample use of it, as they do, for example, in Gershman et al. (2016).

The different computational expansions of ToM models correspond to potentially different neural implementations of the detection of automaticity. Thinking about another person's thoughts or feelings recruits a systematic set of brain regions collectively termed the mentalizing network, or the ToM network (Koster-Hale & Saxe, 2013; Saxe & Powell, 2006; Schurz et al., 2014; Spunt et al., 2015). Debates continue about the specific selectivity of this network, and whether it is better seen as a bona fide ToM network, or instead a social-brain network, or perhaps more related to narrative comprehension (Lin et al., 2018). However, mostly absent from these important debates is the question of detecting and reasoning about scripted behavior in others. It would be interesting to use neuroimaging techniques to first identify ToM networks in individual participants and then test how these brain regions respond to stimuli that vary in the degree to which the person in the stimuli seems to be behaving automatically, either in a naturalistic bottom-up fashion or in a top-down-directed fashion. A priori, it is possible that such reasoning is part and parcel of interacting with others, in which case seeing others behaving automatically would still reliably trigger the ToM network (and perhaps specific subparts within in). However, it is also possible that perceiving automatic behavior in others does not activate the ToM network and activates instead self-action regions related to model-free behavior or that it is processed in a way unrelated to both self-action and ToM and more similar to reasoning about objects.

At the risk of repetition, we again caution that our results need to be further tested for generalization along several dimensions. Our video clips were curated to match many of the interactions people experience, but these mostly involved people performing "for" the camera (including in cases like vlogs and commentaries). It is in principle possible that the lack of a relationship between automaticity and other examined attributes (or the weak relationship with informativeness) may not hold for everyday interactions. We note that we do think such situations, to the degree they involve automaticity, are also often *performative* and the results would hold, but further work is needed here. In addition, our results rely on a specific sample of the population and that while the sample is relatively representative of the U.S. population, it may not be representative of varied cultures. We also fully anticipate that the *specific* scripts that people use and the situations in which they use them will vary across time, development, and cultures. In particular, it may be the case that "tighter" cultures, which are more resistant to change and place more of an emphasis on norms and rule following (Gelfand, 2019), would involve more reasoning about automaticity. However, one could argue a priori that "looser" cultures, which rely less on social cohesion and coordination, would require less mental variable tracking. While empirical work is needed to examine potential differences between cultures, our own current expectation is that while specific scripts may vary, the overall recognition that other people are behaving automatically (and downstream consequences) is robust and relies on overall similar reasoning.[3]

All the world's a stage, and all the men and women merely players. However, beyond the specific lines our fellow actors say, it matters to us if the show is improv or Ibsen. The perception that other people are acting in a scripted way is an important aspect of our daily lives but one so routine that we often sleepwalk through it. It is only when things go awry that we notice we had been expecting something else all along. Like a missed last step in a staircase in the dark, our mind tumbles when what was supposed to be an unstudied moment turns mechanic but also when the very mindlessness of the situation is suddenly covered in, uh, doubt.

[3] A colleague from Germany has on several occasions expressed astonishment that customer service workers in the United States do not *actually* want to know how he is doing, wondering "why do they ask, then?" It is doubtful that Germany does not have scripted behavior, while the United States does, and more likely that their respective scripts vary.

References

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 914–930. https://doi.org/10.1037/0278-7393.33.5.914

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), Article 0064. https://doi.org/10.1038/s41562-017-0064

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*, 48–69. https://doi.org/10.1038/npp.2009.131

Bass, I., Espinoza, C., Bonawitz, E., & Ullman, T. (2024). Teaching without thinking: Negative evaluations of rote pedagogy. *Cognitive Science*, *48*(6), Article e13470. https://doi.org/10.1111/cogs.13470

Berke, M., Tenenbaum, A., Sterling, B., & Jara-Ettinger, J. (2023). *Thinking about thinking as rational computation*. PsyArXiv. https://doi.org/10.31234/osf.io/e65p3

Bonawitz, E., & Shafto, P. (2016). Computational models of development, social influences. *Current Opinion in Behavioral Sciences*, *7*, 95–100. https://doi.org/10.1016/j.cobeha.2015.12.008

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, *22*(6), 956–962. https://doi.org/10.1016/j.conb.2012.05.008

Collins, A. G., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576–586. https://doi.org/10.1038/s41583-020-0355-6

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, *43*, Article e28. https://doi.org/10.1017/S0140525X19001730

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(45), 13817–13822. https://doi.org/10.1073/pnas.1506367112

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*, 429–453. https://doi.org/10.3758/CABN.8.4.429

De Freitas, J., & Alvarez, G. A. (2018). Your visual system provides all the information you need to make moral judgments about generic visual events. *Cognition*, *178*, 133–146. https://doi.org/10.1016/j.cognition.2018.05.017

De Pinto, M. G., Polignano, M., Lops, P., & Semeraro, G. (2020). Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. *2020 IEEE conference on evolving and adaptive intelligent systems (EAIS)* (pp. 1–5). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/EAIS48028.2020.9122698

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *308*(1135), 67–78. https://doi.org/10.1098/rstb.1985.0010

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Doya, K., Samejima, K., Katagiri, K., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Computation*, *14*(6), 1347–1369. https://doi.org/10.1162/089976602753712972

Evans, O., Stuhlmuller, A., & Goodman, N. D. (2016). Learning the preferences of ignorant, inconsistent agents. *30th AAAI conference on artificial intelligence* (pp. 323–329). Association for the Advancement of Artificial Intelligence. https://doi.org/10.1609/aaai.v30i1.10010

Fox Tree, J. E. (2002). Interpreting pauses and ums at turn exchanges. *Discourse Processes*, *34*(1), 37–55. https://doi.org/10.1207/S15326950DP3401_2

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, *8*(9), 396–403. https://doi.org/10.1016/j.tics.2004.07.002

Gao, T., Baker, C. L., Tang, N., Xu, H., & Tenenbaum, J. B. (2019). The cognitive architecture of perceived animacy: Intention, attention, and memory. *Cognitive Science*, *43*(8), Article e12775. https://doi.org/10.1111/cogs.12775

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: Interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 669–684. https://doi.org/10.1037/a0020735

Gelfand, M. (2019). *Rule makers, rule breakers: Tight and loose cultures and the secret signals that direct our lives*. Scribner.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193. https://doi.org/10.1016/0010-0277(95)00661-h

Gershman, S. J., Gerstenberg, T., Baker, C. L., & Cushman, F. A. (2016). Plans, habits, and theory of mind. *PLOS ONE*, *11*(9), Article e0162246. https://doi.org/10.1371/journal.pone.0162246

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. https://doi.org/10.1126/science.aac6076

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*(1), 182–194. https://doi.org/10.1037/a0030844

Goldwater, M. B., Gershman, S. J., Moul, C., Ludowici, C., Burton, A., Killer, B., Kuhnert, R.-L., & Ridgway, K. (2020). Children's understanding of habitual behaviour. *Developmental Science*, *23*(5), Article e12951. https://doi.org/10.1111/desc.12951

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 3909–3917). Curran Associates, Inc. https://doi.org/10.48550/arXiv.1606.03137

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hawkins, R., Goodman, N., & Goldstone, R. (2019). The emergence of social norms and conventions. *Trends in Cognitive Sciences*, *23*(2), 158–169. https://doi.org/10.1016/j.tics.2018.11.003

Hawkins, R., Gweon, H., & Goodman, N. (2021). The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive Science*, *45*(3), Article e12926. https://doi.org/10.1111/cogs.12926

Heidegger, M. (1967). *Being and time* (J. Macquarrie & E. Robinson, Trans.). Blackwell.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259. https://doi.org/10.2307/1416950

Heller, D., Arnold, J. E., Klein, N. M., & Tanenhaus, M. K. (2015). Inferring difficulty: Flexibility in the real-time processing of disfluency. *Language and Speech*, *58*(2), 190–203. https://doi.org/10.1177/0023830914528107

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110. https://doi.org/10.1016/j.cobeha.2019.04.010

Jara-Ettinger, J., & Dunham, Y. (2024). *The institutional stance*. PsyArxiv. https://doi.org/10.31234/osf.io/pefsx

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. https://doi.org/10.1016/j.tics.2016.05.011

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64. https://doi.org/10.1016/j.cognition.2017.06.017

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(45), 12868–12873. https://doi.org/10.1073/pnas.1609094113

Kidd, C., White, K. S., & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science*, *14*(4), 925–934. https://doi.org/10.1111/j.1467-7687.2011.01049.x

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, *13*(4), 400–408. https://doi.org/10.1093/cercor/13.4.400

Kleiman-Weiner, M. (2018). *Computational foundations of human social intelligence* [Doctoral dissertation, Massachusetts Institute of Technology]. DSpace@MIT. https://dspace.mit.edu/handle/1721.1/120621

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron*, *79*(5), 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Kryven, M., Ullman, T. D., Cowan, W., & Tenenbaum, J. B. (2021). Plans or outcomes: How do we attribute intelligence to others? *Cognitive Science*, *45*(9), Article e13041. https://doi.org/10.1111/cogs.13041

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science*, *14*(5), 402–408. https://doi.org/10.1111/1467-9280.01454

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, Article e1. https://doi.org/10.1017/S0140525X1900061X

Liljeholm, M., Tricomi, E., O'Doherty, J. P., & Balleine, B. W. (2011). Neural correlates of instrumental contingency learning: Differential effects of action–reward conjunction and disjunction. *Journal of Neuroscience*, *31*(7), 2474–2480. https://doi.org/10.1523/JNEUROSCI.3354-10.2011

Lin, N., Yang, X., Li, J., Wang, S., Hua, H., Ma, Y., & Li, X. (2018). Neural correlates of three cognitive processes involved in theory of mind and discourse comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *18*(2), 273–283. https://doi.org/10.3758/s13415-018-0568-6

Loy, J. E., Rohde, H., & Corley, M. (2017). Effects of disfluency in online interpretation of deception. *Cognitive Science*, *41*(S6), 1434–1456. https://doi.org/10.1111/cogs.12378

Marticorena, D. C., Ruiz, A. M., Mukerji, C., Goddu, A., & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental Science*, *14*(6), 1406–1416. https://doi.org/10.1111/j.1467-7687.2011.01085.x

Mazar, A., & Wood, W. (2022). Illusory feelings, elusive habits: People overlook habits in explanations of behavior. *Psychological Science*, *33*(4), 563–578. https://doi.org/10.1177/09567976211045345

McCoy, J. P., & Ullman, T. D. (2018). A minimal turing test. *Journal of Experimental Social Psychology*, *79*, 1–8. https://doi.org/10.1016/j.jesp.2018.05.007

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford University Press. https://doi.org/10.1093/0198236107.001.0001

Orena, A. J., & White, K. S. (2015). I forget what that's called! Children's online processing of disfluencies depends on speaker knowledge. *Child Development*, *86*(6), 1701–1709. https://doi.org/10.1111/cdev.12421

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, *24*(5), 751–761. https://doi.org/10.1177/0956797612463080

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. *Proceedings of the 35th international conference on machine learning* (pp. 4218–4227). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v80/rabinowitz18a.html

Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition*, *224*, Article 105073. https://doi.org/10.1016/j.cognition.2022.105073

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, *2*(9), 661–670. https://doi.org/10.1038/35090060

Rutherford, M. D., & Kuhlmeier, V. A. (Eds.). (2013). *Social perception: Detection and interpretation of animacy, agency, and intention*. MIT Press. https://doi.org/10.7551/mitpress/9780262019279.001.0001

Sartre, J. (1956). *Being and nothingness*. Random House.

Saxe, R. (2012). The happiness of the fish: Evidence for a common theory of one's own and others' actions. In M. D. Rutherford & V. A. Kuhlmeier (Eds.), *Handbook of imagination and mental simulation* (pp. 257–309). Psychology Press.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, *17*(8), 692–699. https://doi.org/10.1111/j.1467-9280.2006.01768.x

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum.

Schmidt, M. F., & Rakoczy, H. (2023). Children's acquisition and application of norms. *Annual Review of Developmental Psychology*, *5*(1), 193–215. https://doi.org/10.1146/annurev-devpsych-120621-034731

Scholl, B. J., & Gao, T. (2013). Perceiving animacy and intentionality: Visual processing or higher-level judgment. In M. D. Rutherford & V. A. Kuhlmeier (Eds.), *Social perception: Detection and interpretation of animacy, agency, and intention* (pp. 197–229). MIT Press. https://doi.org/10.7551/mitpress/9780262019279.001.0001

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309. https://doi.org/10.1016/S1364-6613(00)01506-0

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Shu, T., Bhandwaldar, A., Gan, C., Smith, K., Liu, S., Gutfreund, D., Spelke, E., Tenenbaum, J. B., & Ullman, T. (2021). Agent: A benchmark for core psychological reasoning. *Proceedings of the 38th international conference on machine learning* (pp. 9614–9625). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v139/shu21a/shu21a.pdf

Spelke, E. S. (2022). *What babies know: Core knowledge and composition* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780190618247.001.0001

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Spunt, R. P., Meyer, M. L., & Lieberman, M. D. (2015). The default mode of human brain function primes the intentional stance. *Journal of Cognitive Neuroscience*, *27*(6), 1116–1124. https://doi.org/10.1162/jocn_a_00785

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press. https://doi.org/10.1109/TNN.1998.712192

Tanaka, S. C., Balleine, B. W., & O'Doherty, J. P. (2008). Calculating consequences: Brain systems that encode the causal effects of actions. *Journal of Neuroscience*, *28*(26), 6750–6755. https://doi.org/10.1523/JNEUROSCI.1808-08.2008

Templeton, E. M., Chang, L. J., Reynolds, E. A., Cone LeBeaumont, M. D., & Wheatley, T. (2022). Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4), Article e2116915119. https://doi.org/10.1073/pnas.2116915119

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Transaction Publishers. https://doi.org/10.5962/bhl.title.55072

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(34), 8491–8498. https://doi.org/10.1073/pnas.1804761115

Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception & Psychophysics*, *68*, 1047–1058. https://doi.org/10.3758/BF03193364

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems* (Vol. 22, pp. 1874–1882). Curran Associates. https://dspace.mit.edu/handle/1721.1/61347

Vallortigara, G. (2012). Aristotle and the chicken: Animacy and the origins of beliefs. In A. Fasolo (Ed.), *The theory of evolution and its impact* (pp. 189–199). Springer. https://doi.org/10.1007/978-88-470-1974-4

Williams, J., Kleinegesse, S., Comanescu, R., & Radu, O. (2018). Recognizing emotions in video using multimodal DNN feature fusion. *Proceedings of grand challenge and workshop on human multimodal language (challenge-HML)* (pp. 11–19). Association for Computational Linguistics. https://aclanthology.org/W18-3302/

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), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron*, *75*(3), 418–424. https://doi.org/10.1016/j.neuron.2012.03.042

Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press. https://doi.org/10.7551/mitpress/8441.001.0001

Zhang, C., Kemp, C., & Lipovetzky, N. (2023). Goal recognition with timing information. *Proceedings of the International Conference on Automated Planning and Scheduling*, *33*(1), 443–451. https://doi.org/10.1609/icaps.v33i1.27224

Received May 30, 2024
Revision received August 22, 2024
Accepted August 24, 2024 ■