Puzzles of Human Behavior: Revealing Intentional Action Understanding in Typically Developing and Autistic Individuals

By

Joanna Korman

B.A., Williams College, 2007

M.Phil., Cambridge University, 2008

Sc.M., Brown University, 2011

Dissertation

Submitted in partial fulfillment of the requirements for the Degree of

Doctor of Philosophy in the Department of Cognitive, Linguistic,

and Psychological Sciences at Brown University

PROVIDENCE, RHODE ISLAND

MAY 2017

This dissertation by Joanna Korman is accepted in its present form by the Department of Cognitive, Linguistic, and Psychological Sciences as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____
Dr. Bertram Malle, Advisor

Recommended to the Graduate Council

Date_____ _____
Dr. Dave Sobel, Reader

Date_____ _____
Dr. Steven Sloman, Reader

Date_____ _____
Dr. Fiery Cushman, Reader

Approved by the Graduate Council

Date_____ _____
Dr. Andrew Campbell,
Dean of the Graduate School

JOANNA KORMAN

Curriculum Vitae

Born: July 1, 1985, Cleveland, OH
Joanna_Korman@brown.edu
Department of Cognitive, Linguistic, & Psychological Sciences
Brown University
190 Thayer St., Box 1821
Providence, RI 02912

EMPLOYMENT

National Research Council Postdoctoral Research Associate, United States Naval Research Laboratory (NRL), Intelligent Systems Division, Washington DC — *Beginning Oct. 2016*

EDUCATION

Brown University (Providence, RI) — Expected, September 2016
PhD Candidate, Experimental Social Psychology

University of Cambridge (Cambridge, UK) — 2008
MPhil, History and Philosophy of Science *(with distinction)*

Williams College (Williamstown, MA) — 2007
B.A. Psychology (highest honors), and Cognitive Science (honors)
magna cum laude

AWARDS AND HONORS

National Academy of Sciences / National Research Council — 2016-
Postdoctoral Research Associateship Award, total amount up to 3 years: $229,134 (pending yearly review)

American Psychological Foundation William C. Howell Scholarship, $1,000 — 2015

Society for Personality and Social Psychology (SPSP) Travel Award, $500 — 2015

Brown University Dissertation Fellowship — 2015

Brown International Travel Fund Award, $600 — 2014

Brown Brain Science Graduate Research Award (1 semester graduate — 2012

stipend)

Honorable Mention, NSF Graduate Research Fellowship	2010
Brown University Graduate Fellowship, 2009-2010	2009
Cambridge University Overseas Trust Scholar, (*Honorary),* £2500	2007 – 2008
Herchel Smith Fellowship for study at University of Cambridge, $40,000	2007 – 2008
Phi Beta Kappa	Inducted 2007
Sigma Xi	Inducted 2007
Williams College Summer Science Fellowship	2006
Class of 1960's Scholar in Psychology	2005

TALKS

Korman, J. (2016, April). Tearing up the script: Grappling with puzzling behavior in typical development and autism. Invited talk presented at Bard College, Annandale-on-Hudson, NY.

Korman, J. (2016, April). Tearing up the script: The function of reason explanations. Invited talk presented to the Naval Research Laboratory, Intelligent Systems Division. Washington, D.C.

Korman, J. (2016, March). Mechanisms of action understanding in high-functioning autism. Talk presented to the Moral Psychology Research Lab and Moral Cognition Lab, Harvard University.

Korman, J. (2015, November). Inference or integration? Accounting for performance on advanced theory of mind tasks in ASD adults. Talk presented to the Social Cognitive Science Brown Bag, Brown University.

Korman, J. Cusimano, C. & Malle, B.F. (2015, June). Fitting the final puzzle piece: Choosing between belief and desire in intentional action explanation. Talk presented at the Annual Meeting of the Society for Philosophy and Psychology, Durham, NC.

Korman, J. (2014, May). The Rhyme in Reasons: Desire and belief at work in action explanation. Invited talk presented to the Boston Area Moral and Social Cognition group, Boston University.

Korman, J. (2012, November). Walking the walk of mental state talk: Reason explanations in typical development and autism. Invited talk presented to Philosophy Department Colloquium & Midwest Empirical and Theoretical Association (META), University of Illinois-Urbana Champaign.

Korman, J. & Malle, B.F. (2012, June). Practical Rationality in Action Explanation: A crucial role for belief reasons. Talk presented at the Annual Meeting of the Society for Philosophy and Psychology, Boulder, CO.

Korman, J. (2012, May). Walking the walk of mental state talk: The function of reasons in action explanation. Talk presented at the Social Cognitive Science Brown Bag, Brown University.

POSTER PRESENTATIONS

Kim, B. Korman, J., & Malle, B.F. (2016, August). Moral Social Media: Heavy Facebook users accept harsher moral criticism for microaggressions. Poster presented at the annual meeting of the Cognitive Science Society, Philadelphia, PA.

DiFabrizio, B., Korman, J., Cusimano, C., & Malle, B.F. (2016, June). A Hierarchy interrupted? How valence and surprise alter the hierarchy of social inferences. Poster presented at the annual meeting of the Society for Philosophy and Psychology, Austin, TX.

Korman, J., Malle, B.F., Leboyer, M., Gaman, A., & Zalla, T. (2016, May). Inference or Integration? Mechanisms of mental state understanding in high-functioning autism. Poster presented at the Annual International Meeting for Autism Research, Baltimore, MD.

Korman, J., Zalla, T., & Malle, B.F. (2016, January). Inference or Integration? Social cognitive deficits in adults with Autism Spectrum Disorders. Poster presented at the Annual meeting of the Society for Personality and Social Psychology, San Diego, CA.

Korman, J. & Malle, B.F. (2015, February). The rhyme in reasons: Desires as rational explanations for human action. Poster presented at the Annual Meeting of the Society for Personality and Social Psychology, Long Beach, CA.

Korman, J., Cusimano, C. Monroe, A., Smith, J., and Malle, B. F. (2014, July). Not so bad after all? The role of explanation features in blame mitigation. Poster presented at the Annual Meeting of the Cognitive Science Society, Quebec City, Canada.

Korman, J. & Malle, B.F. (2013, January). Keeping mental states in mind: Behavior Explanation in Autism Spectrum Disorders. Poster presented at the Annual Meeting of the Society for Personality and Social Psychology, New Orleans, LA.

Korman, J. & Malle, B.F. (2012, January). No such thing as nonsense: No matter how strange, actions have reasons. Poster presented at the Annual Meeting of the Society for Personality and Social Psychology, San Diego, CA.

PUBLICATIONS

Korman, J. & Malle, B.F. (*in press).* Grasping for traits or reasons? How people grapple with puzzling social behaviors. *Personality and Social Psychology Bulletin.*

Korman, J., Voiklis, J., & Malle, B.F. (2015). The Social Life of Cognition. *Cognition*, 135, 30-35.

Malle, B. F., and Korman, J. (2015). Attribution theory. In B. S. Turner (Ed.), Wiley-

Blackwell Encyclopedia of Social Theory. Malden, MA: Wiley Blackwell.

Korman, J. (2014). Review of *Addiction and Responsibility*. *Philosophical Psychology,* 27, *6,* 930-934.

Malle, B. F., & Korman, J. (2013). Attribution theory. In D. S. Dunn (Ed.), *Oxford Bibliographies in Psychology.* New York: Oxford University Press.

Korman, J. (2011). Concept revision is sensitive to changes in category structure, causal history. *Behavioral and Brain Sciences,* 34, *3,* 135-136.

MANUSCRIPTS IN PREPARATION

Korman, J., Zalla, T. & Malle, B.F. (in preparation). Competence or Performance? Inferential and conceptual deficits in high-functioning autism.

Korman, J. & Malle, B.F. (in preparation a). Walking the walk of mental state talk: Practical rationality in action explanation.

Korman, J. & Malle, B.F. (in preparation b). The function of belief and desire in action explanation.

Korman, J. (in preparation). Lifting the curse of knowledge: The structure of social inference in online behavior observation.

TEACHING EXPERIENCE AND TRAINING

Course Instructor, Introduction to Social and Developmental Psychology, Summer @ Brown, Summers 2013 & 2014

This intensive, 3-hours per day course introduces students to the idea of "theory of mind" and its role in the fields of social and developmental psychology. Precocious summer students gain general facility in reading and critically evaluating articles in the field of experimental psychology, as well as an understanding of the methodological principles and paradigms employed by social and developmental psychologists. Students complete short daily writing assignments, two longer one page papers discussing and critiquing empirical articles, and a final five-page paper synthesizing material from multiple course readings. Through participation in interactive lab activities, students also develop their skills as experimentalists.

Guest Lecturer, Brown University

Social Psychology: "Understanding Other Minds: Categories of Mind," 2016
Leading with Empathy: "Autism and Theory of Mind," 2015
Introductory Psychology: "The Big Five Traits" 2011, 2012
Introductory Psychology: "Social-Cognitive Theory," 2011, 2012
Thinking: "Categorization and Explanation," 2011

Teaching Assistant

Personality and Clinical Assessment (1 term – Brown University)
Introductory Psychology (3 terms – Brown University)
Social Psychology (1 term – Brown University)
Statistics for Psychology (1 term – Williams College)

Pedagogical Training

Sheridan Center for Teaching and Learning, Brown University
Certificate I: Teaching Seminar, 2010-2011
Develop a reflective teaching practice and adapt to distinct learning styles
Certificate III: Professional Development, 2011-2012
Develop a teaching philosophy and hone presentation skills
Certificate V: Principles and Practice in Reflective Mentorship, 2014-2015
Cultivate mentoring skills and participate in undergraduate advising
Team Enhanced Advising and Mentoring (TEAM), 2014-2015
Brown Dean' Office program for discussion of best practices in academic advising

Writing Center Associate, Brown University, 2014 and 2015
One-on-one tutoring of ESL and native-speaker graduate and undergraduate students in argument construction, English language grammar, syntax and usage

MENTORING AND ADVISING

Research Mentorship and Supervision

Baxter DiFabrizio, Brown '15
Co-supervised Brown Undergraduate Training and Research Award, Summer 2014
Co-supervised Psychology Thesis, "The Impact of Script Deviation and Valence on the Hierarchy of Social Inferences," 2014-2015.
Madeline DiGiovanni, Brown '17
Co-supervised Brown Undergraduate Training and Research Award (Summer 2016)
Dominique Moore, Brown '18
Serena Entezary, Boston College '15
Tiffany Phu, Brown '14
Mina Whangbo, Brown '14
Nestor Noyola, Brown '13

Academic and Career Advising

Graduate Mentor, Graduate-Undergraduate Mentoring Initiative, 2014
Mentoring program matching graduate students with first-generation and minority juniors and seniors aspiring to careers in the sciences
Mentee: Luyu Zhang, Brown '14

Primary Academic Adviser to first years and sophomores (2014-2015; 2015-2016)**:**
Aliyah Olinayan, Brown ’18
Joyce Elias, Brown ’18
Meghan Foe, Brown ’18
Dominique Moore ‘18

SERVICE

Social Cognitive Science Brown Bag Coordinator, 2012-13, 2013-14
Departmental Liaison to the Sheridan Center for Teaching and Learning, 2010-2011
Ad Hoc Reviewer:
Cognition
Philosophical Psychology
Cognitive Systems Research

Acknowledgements

Bertram Malle never shies away from hard problems, and he has challenged me to take the same tack. With him I have been able to explore some of the uncharted "open hypothesis spaces" of social cognition that many people will not touch, and he has shared in my relish of the moments when a theory makes itself apparent in such a wilderness. As an advisor, Bertram has been an exceptional match for me intellectually. I was never forced to give up being a philosopher but only to put methodological and mechanistic meat onto abstract theoretical bones. He has also provided ample capital, both monetary and human, to realize these projects, and has instilled in me a love of grant writing along the way.

I also thank my committee. Throughout my time in graduate school, Dave Sobel has provided a crucial developmental perspective on the study of theory of mind. Steve Sloman introduced me to cognitive psychological work on explanation and has provided valuable input on work at various stages. Fiery Cushman has provided valuable and incisive feedback on work at several points in my graduate career.

I am grateful to Tiziana Zalla and her team in Paris at the École Normale Supérieure and at Albert Chenevier Hospital, including Alice Latimier, Marion Leboyer, and Alexandru Gaman. Tiziana has generously lent to this work both her valuable scientific perspective on autism and the resources of her lab and connections with the autism community in Paris.

At Brown I have been blessed with some exceptional colleagues. Andrew Monroe has been my beloved big brother from the beginning. He socialized me into the weird and wonderful world of research, and has been a truly magnanimous friend and mentor throughout my time in graduate school. Corey Cusimano inspired and improved my science with his computational

prowess, and has offered invaluable friendship as well as rare intellectual companionship around fundamental questions in the folk psychology of mind and action. Steve Guglielmo has been an important role model, and has generously offered his keen perspective over many years. Anna Hartley has personified fortitude, pluck, and statistical savvy as long as I've known her. She has walked both ahead of me and beside me in this process, both as a scientist and as a human being. It would have been immeasurably more difficult without her.

In a very short time, Stuti Thapa Magar has become a respected colleague and cherished friend. She has stepped up enormously in the last year to make this work possible, lending powerful statistical insight and programming chops to a number of the projects in this dissertation and a grounding of sisterly comradery to my day-to-day work life. Boyoung Kim has been an energetic, diligent collaborator as well as an exceptionally generous friend. Jie Ren has been the best officemate I could have asked for; Mark Ho has provided a valuable listening ear.

Within the autism community in Rhode Island and greater New England, I have been privileged to come into contact with many generous and resourceful leaders, researchers, and stakeholders. At the Asperger's Association of New England, Toby Liebowitz, Arthur Mercurio, and Max Sederer opened their doors to my early recruitment efforts. The formation of RICART was crucial to this work, and I am grateful to the people who built and sustained this effort. In particular, Steve Scheinkopf, Lindsay Oberman, Alan Gerber, and Alicia Eid all went out of their way to help make this project a success. In addition, meeting and interacting with individuals on the autism spectrum and their families has been a highlight. Without the openness of these individuals themselves to participation in research, none of our work on autism would have been possible.

Thanks also to Warren Bilker and the team at UPenn for their cheerful assistance on our development of our version of the online version of Ravens Progressive Matrices, as well as to Nathalie Oulhen for serving as a reliable and willing translator on the French-language work.

In addition, this dissertation would not have been possible without the dedicated work of a number of superb current and former undergraduate research assistants. As a staff member on the project, Baxter DiFabrizio has brought supple intelligence and boundless enthusiasm. Maddy DiGiovanni and Dominique Moore have devoted significant time and thought to their work as coders, and have each enhanced the project with their excellent questions. Emorie Beck has been truly game for anything, making myriad contributions from flyering to high-level MATLAB wrangling. Serena Entezary, Tiffany Phu, Fue Vue, and Mina Whangbo also made contributions.

My two mentors at Williams College, Ari Solomon and Joe Cruz, each took considerable time and care to nurture my penchant for psychological theory. At the University of Cambridge, Martin Kusch warmly encouraged me to articulate the philosophical power and boundaries of folk psychological categories. At Columbia Psychiatry, Madelyn Gould was an exceptional role model, and she and Alison Lake showed me the power of targeted community-based research with clinical populations.

Many friends from both inside and outside the ivory tower have provided intellectual stimulation, moral support, and distraction, never wavering even when I fell off the map for months and sometimes years at a time: Elysia Alleman, Katy Dieber, Emily Button Kambic, Samreen Kazmi, Michelle Mazala Beth Myre, Lucy Seminoff-Flam, Raphael Shargel, Graeme B. Schranz, Allison Smith, Ellen Weinstein, and Greg Winger. My sister Amanda has fortified

me with her credible and unwavering faith. Thanks also go to my brother Ben, and to friend and renaissance man Brian Hirshman, each of whom has offered invaluable advice.

The unflagging support and generosity of my parents, Neil and Diane, over my entire life and in the past few years, has made the pursuit and completion of this work possible. My mother has modeled fortitude and compassion in her professional life and in her raising of me, and my father has been a significant professional role model and mentor.

My in-laws, Anne and Jim, have cheered me on without fail, expressing curiosity but never nosiness, and have made me feel incredibly welcome in their family. The moral support and warm hospitality of my Uncle Craig and Aunt Carol have also meant a great deal to me during this time.

My husband Tom is the voice of reason and chief comedian of my life. He is a pedantic feminist and warrior against nonsense who loves me with a bold and stubborn persistence. Over the last several years of our commuter marriage across lines of state and of country he has logged tens of thousands of miles of driving in the midst of his own demanding work schedule -- just to be by my side. He has taught me that most of the problems I have faced are, if not solvable, at least ones that I can count in common with the inestimable Lisa Simpson. His support of my aspirations and of my sanity in the face of many steep challenges over these long, lean years has been little short of heroic.

My beautiful, loving, and mild-tempered cat, Yitze, surpasses everyone in raw number of hours of companionship offered. She is a reminder to all of us that no matter how remote the possibility, we should hold out for peace in the middle east.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1:

Introduction

1.1. The Concept of Intentional Action

Adult social perceivers understand the meaning of a behavior using the concept of *intentional action*. People consider an action to be performed *intentionally* only if that action has a certain set of features: The agent needs to have had a *desire* for the action's outcome, a *belief* that the action would lead to that outcome, and an *intention* to bring about that outcome right before doing so (Kashima, McKintyre, & Clifford, 1998; Malle & Knobe, 1997). Social perceivers assume that an agent's beliefs and desires play a causal role in bringing about her intentional action: they are the mental states *in light of which* that agent is thought to have performed that particular action. Additionally, these mental states have a rational component. When an observer of action generates mental states to explain an agent's behavior, she also understands the content of those mental states as clarifying how the agent's *reasoning* brought about the particular action in question. In virtue of their rational and causal aspects, mental states acquire the status of reasons (Davidson, 1963; Malle, 2004, 2011). How do reasons help people to puzzle out the precise meaning of an action? Because intentional action is always "action under a description," an observed action (e.g., a man operating a water pump) may lend itself to multiple interpretations depending on the reasons for which the agent acted: while some effects of the action were *intentional*, other effects may not have been (Anscombe, 1965, Section VI, p. 11-12). Depending on his mental states, the man may be merely "pumping water to the house below" (if he *knows* the supply reaches that house), or he may actually be "poisoning Nazis," if in addition, he *knows* the water to be poison and *knows* there are Nazis living in the

house below. The observer's understanding of the agent's reasoning process thus determines how she resolves this ambiguity.

This dissertation explores the limits of application of social perceivers' use of the concept of intentional action. Specifically, it addresses whether or not the use of this concept extends to social perceivers' understanding of two distinct types of behavioral puzzles. Before discussing these two types of puzzles in greater depth, I provide a brief sketch of key developmental achievements in the understanding of intentional action.

1.2 Intentional Action: Development of Component Concepts

Because it is comprised of multiple components, the concept of intentional action does not develop all at once. Instead, a full understanding of the varied causal interactions between subcomponents – and thus, sensitivity to inconsistencies among those components -- emerges only after each of these subcomponents, such as desire and belief, are already in place. In this section, I provide brief snapshots of developmental points that provide crucial conceptual building blocks for the typically developing child's eventual full grasp of the concept of intentional action.

Long before they are capable of explicit inferences about a person's subjective desires, infants are able to detect the goal of an observable, goal-directed action by otherwise inanimate objects (Gergely & Csibra, 1997) as well as by human agents (Woodward, 1998). Infants are especially sensitive to inanimate objects' performances of goal-directed action when the appropriate abstract cues are present, such as being self-propelled or resulting in a salient outcome; they also infer goals after viewing repeated, equifinal variations of the same action (Biro & Leslie, 2007). In addition, infants show sensitivity to the relationship between goals and

means, using information about the goal of an action to predict the most efficient means to achieve that goal (Gergely, Nádasdy, Csibra, & Bíró, 1995). They also use information about the means of an action to select from multiple candidate goals (Verschoor & Biro, 2012). The infant's use of means in determining her inferences about the structure of action may support her ability to recognize hierarchies of goals in action, and to see the relevance of actions (or lack thereof) to one another within a sequence.

Between the ages of one and three, children make one of the first transitions from understanding behaviors to understanding minds: they move from detecting the *observable* features of goal-directed behavior, to understanding that others have desires: internal, unobservable psychological states that are consistent with (Wellman & Woolley, 1990) and may cause (Repacholi & Gopnik, 1997) their actions and emotions. Whereas earlier in development, all experimental measures of action understanding must be implicit (e.g., Gergely et al., 1995), children articulate desires in explicit language, referring to them in their explanations of action from about the age of two (Bartsch & Wellman, 1995). Belief understanding follows. Around age 3, children grasp the need to articulate their own and others' beliefs (which are "invisible" representations to which others may not have access); only at age 4 do they then do they begin to make an explicit causal connection between these states and people's intentional actions, providing beliefs in their explanations for action (Bartsch & Wellman, 1995). Passage of explicit false belief tasks, in which children must correctly predict a character's action — where she will look for an object hidden in her absence — by attributing a false belief to her, similarly does not typically occur until the age of four (Wellman, Cross, & Watson, 2001).

Although many of the main components of intentional action understanding are in place by age four, the explicit understanding of the subtleties of how beliefs, desires, and intentions

relate to actions as they play out in the world is still emerging. For example, intention and desire are separate concepts. While intentions are representations of an agent's planned future actions (Baird & Astington, 2005), desires are merely representations of desired outcomes or states of affairs that exist regardless of action. For example, a person may generally like candy. Should this person open her drawer in search of a pen, only to find a Snickers bar, this person's general desire for a Snickers bar would be generally fulfilled, but it would not be fulfilled intentionally. In contrast, a Snickers bar may be obtained intentionally if she walks to a drugstore with the specific intention to purchase one, and does so.

Even more complex, not all of an agent's desires are necessarily relevant to the intentional actions he carries out. Particular desires are linked to (and serve to explain) intentions only when they are specifically on the agent's mind right before the agent performs an action in the service of the fulfillment of that intention. And it is even possible to have general desires that may appear to *conflict* with an intended action. As Feinfield, Lee, Flavell, Green, & Flavell (1999) describe, a child may himself desire to go to the park, but if his mother wants him to go to school instead, he may form an intention to go to school, thus fulfilling, with this particular action, a different (and conflicting) desire: to please his mother. Furthermore, if he makes a wrong turn on his way to school and ends up at the park by accident, he has fulfilled his earlier desire to go to the park, but he has not fulfilled the intention to go to school, nor has he fulfilled the desire that led him to form that intention: to please his mother. In other words, although the outcome matches his original desire, it does not match his intention. In order to understand that the intention has been fulfilled, the social perceiver cannot merely compare *any* of the agent's desires to the outcome: she must, in this case, know that the *relevant* desire is the one that plays a causal role in the formation of the intention itself. Feinfield et al. (1999) as well

as Schult (1996, 2002) demonstrate that while three-year-olds struggle to differentiate desire and intention in such scenarios, children between the ages of 4 and 5 have come to understand that an outcome that fulfills a person's general desire does not necessarily fulfill his specific intention (although see Phillips, Baron-Cohen, & Rutter, 1998 for an exception in the case of perceiving one's *own* desires versus intentions). Baird & Moses, (2001) demonstrate further development, showing that five but not four-year-olds can recognize that concrete, observable actions are conceptually distinct from the intentions they fulfill, and that the very same action can thus be motivated by different intentions for different agents. Although development of intentional action understanding proceeds beyond this point and incorporates additional components, such as the agent's skill to perform the behavior in question, the main components of interest in this dissertation – belief, desire, intention, and their causal connections to action – are in place in typically developing children by age 5.

1.3 Intentional Action and Beyond: Repairing Breakdowns in Action Understanding

Both in response to explicit prompts and when observing intentional behavior in everyday contexts, adults' competent use of the concept of intentional action is often guided by scripts (Schank & Abelson, 1977; Karniol, 2003) that limit the information they draw upon. In these cases, the unobservable intention underlying a behavior matches the described action's most obvious meaning. For example, if Joe is at the cash register of the convenience store buying milk and eggs, and has the prior intention *to hand the cashier five dollars*, and he successfully *hands the cashier five dollars,* then no other contents require consideration: the prior intention matches the action, and there are no unintended outcomes to account for. This interpretation of Joe's action assumes, and does not need to explicitly consider, that *Joe knew his*

milk and eggs cost five dollars, and that Joe believed that *if he handed the cashier five dollars, the food would be paid for,* and that *Joe wanted to pay for his groceries.* Because these normally unobservable components come "for free" when an action and its underlying intention clearly cohere, coming to a correct understanding of the behavior is simple for the typically developing social perceiver. Indeed, understanding intentional behavior is often as simple as describing what another person is doing. For example, "What's she doing?" "She's [intentionally] watering the flowers." Or: "What's he doing?" "He's [intentionally] picking up his daughter at school."

Only when an action is inconsistent with the social perceiver's prior, scripted knowledge, and thus, fails to deliver understanding, do they need to look beyond a behavior's most obvious surface features to explain its meaning. For example, when first observing another agent reaching over to a door handle, the social perceiver may understand very quickly that the agent's goal is *to open the door*. In this case, the perceiver may recruit her prior knowledge about the way that particular door handle works (e.g., turn the handle to the left and gently push). Such knowledge typically provides specific expectations about the goals and means by which particular intentions are generally fulfilled. If the agent turns the knob to the left, no further explanation is needed; the action is understood. If the agent turns the knob to the *right*, however, then this knowledge will cease to be perfectly consistent with the goal of *opening the door*. This inconsistency may elicit the social perceiver's wonderings about *how* the observed action relates to her current understanding of the agent's intention (*if she is trying to open the door, why is she twisting the handle the wrong way?)*

What conceptual tools do social perceivers use to restore understanding under these and similar circumstances, when the normal consistency that exists between the mental states

considered by the agent before acting, the action chosen, and resulting outcomes breaks down? Are the components of intentional action understanding – mental states tied directly to action – always sufficient to repair these breakdowns? Or are there cases in which breakdowns in intentional action understanding also mean breakdowns in the concept of intentional action? Such puzzles could instead invite the use of other concepts not directly linked to the concept of intentional action, such as the background causes of an agent's mental states. Although the concept of intentional action focuses on the mental states on the agent's mind as she is making the decision to act, these mental states are embedded in a larger causal background of the agent's environment, culture, and personality characteristics. All of these background factors may serve as explanations of the agent's behavior, either in addition to, or in lieu of, the agent's mental state reasons themselves. The first chapter of this dissertation, introduced in (a) below, focuses on this first question, exploring the concepts people use to restore understanding when the assumptions of prior knowledge are challenged. The second chapter follows up on this question, asking whether specific subcomponents of concept of intentional action function to restore understanding in response to particular types of violations of prior knowledge. Finally, the third chapter, introduced in (b) below, explores another type of puzzle that can arise from a realized intentional action: when the outcome of that action directly conflicts with the outcome originally intended by the agent. This chapter asks whether the concept of intentional action, which is commonly used by typically developing individuals to resolve such puzzles, also extends beyond use in typically developing social perceivers to social perceivers on the autism spectrum.

1.3.1 Denying Plausibility: Puzzles of Prior Knowledge Violation (Chapters 2 & 3)

The inferences typically developing adults make from human action have been widely studied in the context of narrative text comprehension (McKoon & Ratcliff, 1992; Graesser, Singer, & Trabasso, 1994). Since narratives are explicit communications, the mental representation of that narrative—or situation model—one builds is only as rich as the text and one's knowledge-based inferences from the text allow. According to the text-processing literature, there is a particular subset of inferences that the reader will make during text comprehension when she has no particular processing goal. Because every sentence is understood to serve some communicative purpose, every new piece of information is integrated into the reader's growing understanding of the narrative in order to maintain coherence. Sometimes, discontinuities in the narrative arise. For example, there may be change from one central protagonist to another, his plans, or an abrupt transition from one event in one sentence to a seemingly unrelated one in the next. In order to preserve the coherence of the narrative, readers make inferences based on (1) the information present in the text and (2) the knowledge structures needed to inferentially link sequences of multiple events and behaviors (Zwaan & Radvansky, 1998).

When narratives depict a plausible situation, discontinuities can generally be resolved by filling in existing gaps with one's own general knowledge. For example, in the doorknob example above, the social perceiver can confidently infer, or at least speculate about, the causes of turning the knob the wrong direction on the basis of her knowledge about how that door (or doors in general) function. Based on this knowledge, the social perceiver poses the question, *If she is trying to open the door, why is she twisting the handle the wrong way?* In response to such a question, the perceiver can further rely on her knowledge to generate inferences (and, if verbally prompted to do so, explanations) that account for this discontinuity. Previous work

(e.g., Malle, Knobe, & Nelson, 2007; Malle, 2004; McClure & Hilton, 1998) demonstrates that in these everyday cases, people are able to explain intentional behavior by generating the agent's mental states: the reasons she had on her mind before acting. When recognizing a conflict, for example, between the agent's apparent desire and the outcome of her action (as will also be explored in Chapter 4), a natural way to resolve this is to search one's own knowledge base for plausible connections between the agent's action and the intended outcome. In the above case, it is highly plausible that the agent simply had a wrong belief about the correct direction in which this particular doorknob should be turned, and this belief conveniently resolves the inconsistency between the intended and observed outcomes: The agent was intentionally turning the knob to the left, but she was only doing this because she wrongly thought that it would result in the door opening.

In contrast, when the actions described deviate considerably from accepted social scripts and schemata about common interactions between people, actions, and objects, it is not clear that same conceptual tools will be used to generate inferences and explanations, or if they are, how the contents of such inferences will be generated. For example, imagine the social perceiver hears about someone reaching for a hairbrush and attempting to insert it into the ignition of a car. While this depiction does provide a low-level description of the actions involved, does not deliver understanding: neither the typical function of hairbrushes, nor the typical object involved in starting a car, are involved, and the social perceiver may be at a loss even for knowing what the agent's most basic goal is. When the very knowledge base that provides the content of most of people's everyday mental state attributions is challenged, one hypothesis is that social perceivers will shift their focus away from the mental states the agent took into account when forming her intention. Instead, they may focus on understanding of the broader circumstances –

such as culture or personality – that would lead a person to perform such a strange action. However, if the conceptual structure of intentional action is tied primarily and inextricably to mental state reasons as the causes of action, people may persist in the search for mental states in spite of these challenges. Chapter 2 of this dissertation explores these two hypotheses, and Chapter 3 follows up on this work with an exploration of how belief and desire reasons each make unique contributions to the restoration of action comprehensibility.

1.3.2 Puzzles of Inconsistency: The Case of the Supplanted Outcome (Chapter 4)

More common for social perceivers than deep-seated violations of prior knowledge, however, are encounters with intentional actions in which the agent's original intentions result in unforeseen outcomes. While many actions have unforeseen outcomes – those that the agent did not know would result from her action -- some unforeseen outcomes may be generally desirable (e.g., happening to run into a good friend at a cafe when one was only going to the café with the intention of picking up a morning coffee). While as unintended outcomes they do reflect an important conceptual inflection point, such outcomes are unproblematic for social perceivers as long as they are compatible with the fulfillment of the original intention (e.g., obtaining coffee). To be sure, the curious social perceiver may search for the causes of these additional outcomes, but the fundamental expectation – fulfillment of the agent's original intention – has still been met.

In other cases, however, unforeseen outcomes may actually *supplant* the original (intended) outcome, precluding its fulfillment in that same action. For example, one can intend to obtain a morning coffee, and fulfill this intention with a trip to the café, *and* have a serendipitous encounter with a friend at that café. In contrast, if one intends to obtain an ice

cream cone and seeks to fulfill that intention by travelling to a certain address where (one expects) the ice cream shop is located, only to discover that this address is now occupied by an abandoned building, the fulfillment of the intention has been precluded: the building is either the ice cream shop or it is not, and if the latter is the case, then the original intention will remain unfulfilled by the planned action (even if one still has the good fortune of running into a friend outside the building). A unique puzzle is thus created when an outcome that *does* result from an action seems to directly conflict with the intended, expected outcome.

In order to resolve this discontinuity, the social perceiver must first detect it: he must apply his existing knowledge to the particular situation presented, inferring that people generally plan a trip to the ice cream shop with the desire to *obtain an ice cream cone,* not to *find an abandoned building.* If the typically developing social perceiver recognizes this knowledge-based discontinuity, he will recognize that that this outcome (encountering an abandoned building) also represents a *conceptual* discontinuity in the fulfillment of this particular actor's intention (encountering an abandoned building does not fulfill his original intention). The social perceiver will then attempt to resolve this discontinuity by making one of a small subset of inferences to answer the question, "If he wants to go to the ice cream shop, why did he go to an abandoned building?" For example, depending on how much background about the protagonist and the ice cream shop is provided in the narrative, the social perceiver might now infer that the protagonist made a wrong turn on his way to the shop (he was mistaken about its actual location), that he is absent-minded and forgot that the shop is located a few blocks away, or that he simply had outdated knowledge (a false belief) about the identity of the building in that location.

Typically developing adults are capable of seamlessly integrating their social knowledge with conceptual understanding in order to solve puzzles of intentional action in this manner. In contrast, the mechanisms underlying intentional action understanding in high-functioning autism are much less well understood. Members of this population, who develop a basic understanding of belief only in early adolescence (Happé, 1995; Tager-Flusberg & Joseph, 2005), present an informative case study in intentional action understanding because even in adulthood, they struggle to correctly resolve narratively presented behavioral puzzles involving intentional actions (Zalla, Sav, Stopin, Ahade, & Leboyer, 2009; Zalla & Leboyer, 2011; Happé, 1994).

While these behavioral puzzles are typically described as "advanced theory of mind tasks," in these tasks heavy emphasis is placed on ASD adults' isolated belief understanding, rather than their understanding of intentional action per se. However, resolving puzzles of intentional action -- such as the puzzle of the supplanted outcome -- requires abilities beyond belief understanding. Specifically, the observer must have both a capacity to use her social knowledge to correctly infer mental state information, as well as a firm grasp of the concept of intentional action: not only what the constituent conceptual components are, but the myriad causal relations among them. Chapter 4 of this dissertation reports on two studies exploring the role of the inference and integration of mental state information for the resolution of puzzles of intentional action in high-functioning autism.

CHAPTER 2:

Grasping for Traits or Reasons? How People Grapple with Puzzling Social Behaviors[1]

2.1 Introduction

A long tradition of social psychological research has emphasized the critical role of stable traits in people's interpretation of others' behavior (Jones & Davis, 1965; Ross & Nisbett, 1991); see Moskowitz & Olcaysoy Okten, 2016 for a review). According to this tradition, trait attribution is the fundamental tool for understanding others' behavior partly because it affords social perceivers efficient and predictively useful knowledge about others (Hastorf, Schneider, & Polefka, 1970; Shaver, 1975). For example, if Jane is seen helping an old woman cross the street, the social perceiver infers that Jane is a *kind* person, or if she works late into the night on her homework, the social perceiver infers that she is *conscientious.* Both of these concise descriptions imply Jane's future behavior—other kind and conscientious acts she may perform.

In the past twenty years, however, a different consensus has emerged, particularly outside of social psychology. As originally suggested by Heider (1958), people indeed focus their explanations of everyday behavior on the *person*. But rather than stable traits, it is a person's mental states — beliefs, goals, and intentions — that constitute the "default" mode of understanding behavior (Malle & Holbrook, 2012; Moskowitz & Olcaysoy Okten, 2016). When people explain why a person performed a particular action, their appeal to mental states integrates aspects of the person (e.g., a goal in performing the action) as well as aspects of the situation (what the goal aims to achieve in the world). People make these kinds of mental state inferences spontaneously (Hassin, Aarts, & Ferguson, 2005), and they do so more easily and quickly than they are able to make trait inferences (Malle & Holbrook, 2012; Van Overwalle,

[1] A version of the current paper by the same title is in press at *Personality and Social Psychology Bulletin*.

Van Duynslaeger, Coomans, & Timmermans, 2012). In their verbal explanations of behavior, too, people primarily infer mental states, not stable traits (Malle, Knobe, & Nelson, 2007; McClure, 2002).

This new consensus crosses numerous disciplines. For half a century, philosophy of action has examined how intentional action is explained by subjective mental states (the agent's reasons) that are at the same time objective causes in the world (Sandis, 2009). Developmental psychology has offered strong evidence for infants inferring goals and preschoolers inferring a variety of other mental states (Bartsch & Wellman, 1995; Gergely, Nádasdy, Csibra, & Bíró, 1995), whereas trait inferences appear later in development (Kalish & Shiverick, 2004; Snodgrass, 1976). Finally, social neuroscience has revealed a network of brain regions specialized for mental state inferences, activated in response to observing human actions (Saxe, Carey, & Kanwisher, 2004).

The "mentalizing" consensus view is so widely held by researchers outside traditional social psychology that some even consider it a foregone conclusion that all behaviors elicit mental states: "Explicit explanations of *any* behavior . . . contain mental state reasons." (Young & Saxe, 2009, p. 1404; emphasis added). But this conclusion is almost certainly overstated, as much research in social psychology shows people's readiness to infer traits (Uleman, Saribay, & Gonzalez, 2008), and people clearly don't explain behavior solely with mental states (Malle, 2011). Furthermore, the mentalizing consensus relies on evidence grounded in a fairly limited scope of behaviors, namely ordinary, expectancy-consistent behaviors that are relatively easy to explain with common beliefs and goals. For example, Van Overwalle et al. (2012) and Malle and Holbrook (2012) — both demonstrating that people make goal inferences more quickly than trait inferences —used ordinary behaviors such as "After paying the bill, she left 5 euros on the

table," and "The woman sweeps the floor in the apartment hallway." Similarly, work on people's free-response explanations has typically presented expectancy-consistent behaviors, such as, "Fred, who frequently went to expensive restaurants, went out for a meal at an expensive restaurant with his brother" (McClure & Hilton, 1998) or "Why did Anne invite Ben for dinner?" (Malle, Knobe, O'Laughlin, Pearce, & Nelson, 2000). These studies confirm that goals and mental states are frequently used for such everyday behaviors. However, the types of behaviors that are most likely to invite why questions or elicit explanatory activity are those that deviate from expectations (e.g., Wong & Wiener, 1981, Wiener, 1985), and there is little evidence that the dominant use of goals and mental states extends to such expectancy-violating behaviors. In fact, Ratcliffe (2007) suggests that the apparent prevalence of mental states as explanations for everyday behavior is due to easy accessibility of scripts and shared norms that imply those mental states. Therefore, when people explain *puzzling* behaviors (ones that violate scripts or norms), they will refer to traits and other causal background factors such as upbringing, habits, and social background — not mental states.

For just these puzzling, expectancy-violating behaviors, theories in the attribution tradition specifically predict that social perceivers are most likely to ascribe traits to an actor—because such behaviors violate "consensus" (Kelley, 1967) or "desirability" (Jones & Davis, 1965) and thus are diagnostic of the agent's unique characteristics (Reeder & Brewer, 1979; Sanbonmatsu, Mazur, Behrends, & Moore, 2015; Skowronski & Carlston, 1989).[2] To illustrate, an extreme violation such as robbing a bank is highly diagnostic of a trait like *dishonesty*, while a mild violation such as lying about one's age is not as diagnostic. Likewise in the positive

[2] Throughout the paper, we focus on expectancies or priors that social perceivers have about others' behaviors *in general* (e.g., people expect others to be polite or to eat sitting down while at a fancy restaurant). Social perceivers may also have expectancies about a particular individual (e.g., that Jared will behave kindly because he is known to be a kind person). However, if such individual-based expectancies are violated, theories such as Kelley's (1967) predict a decreasing use of traits (the case of "low consistency"). To specifically test the prediction of an increasing use of traits we focus on violated expectancies about people, contexts, and behaviors (the case of "low consensus").

domain, volunteering one hour a day in the homeless shelter violates expectations and is highly diagnostic of a trait like *compassion*, whereas thanking a grocery store cashier for a receipt is quite expected and therefore not as diagnostic of a trait such as *friendliness*. Thus, expectancy-violating behaviors are particularly prone to elicit trait attributions.

While current evidence for a hypothesized predominance of mental states is limited to ordinary behaviors, the evidence for a hypothesized predominance of traits is also limited. Evidence for this hypothesis comes from tasks that do not elicit explanations of behavior but invite participants to form an impression of the actor and to use response options restricted to trait descriptors (e.g., Trafimow, et al., 2005). But trait ascription and behavior explanation are distinct in both function and process (Hilton, Smith, & Kim, 1995; Johnson, Jemmott, & Pettigrew, 1984), and findings from one phenomenon do not necessarily generalize to the other. Whereas an observer's behavior explanation accounts for *why* an actor performed a particular behavior by citing specific characteristics of some person or entity involved in the behavior, trait ascriptions are a form of "belief updating": the observer "learn[s] more about the general characteristics of some person or entity" (Hilton, Smith & Kim, 1995, p. 378). So while traits are dominant in forming an impression of an actor who performs an expectancy-violating behavior, they may or may not be as dominant when it comes to actually explaining the expectancy-violating behavior itself.

When considering the phenomenon of explaining puzzling behaviors, evidence from the mentalizing and trait-ascription hypotheses thus suggest competing predictions. While evidence for the mentalizing hypothesis demonstrates that people explain ordinary behaviors with mental states, the evidence does not currently extend to explanations of *expectancy-violating* behaviors. And while evidence for the trait hypothesis demonstrates the increased use of traits in the process

of trait ascription for expectancy-violating behaviors, the evidence does not currently extend to *explanations* of such behaviors. The goal of the present studies, therefore, is to determine which of these two hypotheses best generalizes to the *explanation* of *expectancy-violating*, truly puzzling behaviors. Alongside the trait vs. mental state hypotheses we will also test a broader variant, according to which traits are only one kind of causal background factor (others being culture, norms, or social context; Malle, 2011), and that together these background factors will trump mental states in explanations of puzzling behaviors (Ratcliffe, 2007).

2.2 Methodological Approach

In four studies we invited participants to provide open-ended explanations of a range of moderately to extremely puzzling behaviors. The studies' use of puzzling behaviors avoids limitations of previous explanation studies in which participants could use scripts and schemas to easily retrieve the agent's goals and beliefs. Instead, by driving participants to the edge of their ordinary explanatory habits, the present studies will reveal whether mental states or traits constitute people's default explanatory tool. Furthermore, the studies' use of open-ended explanations avoids limitations of previous trait ascription studies, which often narrowed potential inferences to a very small, pretested set (Winter & Uleman, 1984). In an open-ended task, participants can select whichever explanations (involving mental states, or traits, or other causes) they feel resolve the puzzle at hand.

For this more even-handed measurement context, we can reformulate the two alternative hypotheses developed above: Hypothesis M predicts that even when trying to explain puzzling behaviors, people will continue to routinely cite the agent's mental states. Hypothesis T predicts that people will increasingly cite explanations involving stable traits, or causal background

factors more generally. In testing these hypotheses, we focus on explanations of intentional actions, which people are most interested in explaining (Malle & Knobe, 1997) and for which both mental state inferences (Buss, 1978; Malle, 1999; McClure, 2002) and trait inferences (Jones & Davis, 1965; Shaver, 1975) have been claimed to be pivotal.

To sharpen the hypothesis tests we draw on a theoretical framework of intentional action explanation that ascribes specific functions to such explanatory constructs as mental states, traits, and other causal background factors (Malle, 2004, 2011). According to this theory, people explain intentional actions by citing the agent's mental states (primarily beliefs and desires) as the *reasons* for which the agent acted—that is, the beliefs and desires in light of which and on the grounds of which the agent decided to act. For example, "Ben invited Sarah to dinner because he thought she liked him." This explanation refers to Ben's own reasons, his subjective mental states — he may well be wrong in thinking that Sarah likes him. The second main type of explanation people use to explain intentional behavior is referred to as a causal history of reason (CHR) explanation. Whereas reasons cite the specific mental states on the agent's mind before performing the action, CHRs cite background factors that may have led up to those reasons – factors literally in the causal history of those reasons. Consider a man who goes to the store. One might explain this action by saying that *he wants to pick up three large turkeys for Thanksgiving* (citing a desire reason). Or one might refer to the background for his reason — the fact that *he has eight sons*. The fact that he has eight sons was not something the agent actively had on his mind but it accounts for why he would want three large turkeys in the first place; thus, the fact lies in the "causal history" of that reason. Such CHR explanations can cite a diverse array of causal background factors, including culture, contexts, and traits. Additional examples are provided in Table 2.1.

Table 2.1. Examples of reason and causal history explanations for the behavior, "The police officer gave the teenager a speeding ticket."

Reasons	The police officer could tell that the teenager was speeding in a school zone.	The police officer didn't want the teenager to think he could get away with speeding.	The police officer needed to meet his quota for tickets that month.
Causal Histories	The police officer was suspicious of teenagers.	The police officer was notoriously strict in his enforcement of posted speed limits.	The police officer was having a bad day [1] and taking his frustration out on the teenager. [2]

We used these theoretically derived distinctions and the validated classification system that assesses them (Malle, 1998) to measure (a) the frequency of inferred mental states (number of reason explanations), (b) the frequency of causal history explanations, and (c) among causal histories, the frequency of trait explanations.

Hypotheses M and T differ in a double comparison: in the relative number of *reasons vs. trait/causal history explanations* in response to *puzzling vs. ordinary behaviors*. We considered several designs for the latter comparison. A within-subject design would have drawn undue attention to the difference between puzzling and ordinary behaviors, making any pattern of results vulnerable to conversational demand accounts. A between-subjects design seemed preferable, but there are notable differences in explanation patterns across varying contexts of ordinary behavior explanations (e.g., spontaneous explanations in conversation versus answers to why questions; experimenter-generated versus participant-generated behaviors; Malle, 2006). To maximize generalizability across these and other variations, we took advantage of the substantial number of studies previously conducted that assessed explanations of ordinary (non-puzzling)

behaviors—in particular, six studies from Malle et al. (2007). In the aggregate, these studies represent the currently most reliable comparison standard of ordinary-behavior explanations, with a total sample size of $N = 732$. For example, together with a sample size of $N = 70$ in Study 1, it enabled us to detect effect sizes of $d \geq 0.35$ at $\beta = .8$ and $\alpha = .05$. Statistical power for all other studies is reported in the appendix.

2.3 Study 1

2.3.1 Method

Stimulus Development. Actions may be puzzling in a variety of ways—they may be novel, misfit their context, or violate statistical, social, or moral expectations. We constructed our behaviors to be puzzling with respect to social perceivers' prior knowledge and expectancies about behavior in general, not about the behavior of any single individual. We all but eliminated individual-based expectancies because social perceivers stood in a zero-acquaintance situation with the (fictitious) target agents. Further, we avoided stimuli that strongly violate moral expectations because they would introduce a potential confound, given the legal and everyday importance of mens rea for the moral domain (Cushman, 2008). Among morally largely neutral actions, we represented the variety of puzzling behaviors using four stimulus categories. These categories of puzzles are based on the idea that social scripts (Schank & Abelson, 1977) ground knowledge about the causal and temporal structure of behavioral events and that schemata (Rumelhart & Ortony, 1977) provide the types of agents and objects that are normally co-involved in such events. The first three categories of puzzling behavior derive from cases that break either a script or a schema or both a script and a schema.

Script breaking, schema-compliant. The first category of stimulus sentences had elements that were semantically associated (schema compliant) but violated a familiar script. We created such script violations by altering the expected structure of events (actions denoted by verbs). For example, in the sentence, "The garbage men dropped off bags of trash at the end of each driveway," the verb "drop off," though semantically compatible with the job of garbage men, violates the sequence of actions normally completed by garbage men. This sequence violation was achieved by replacing the phrase "pick up" with the phrase "drop off."

Schema-breaking, script-compliant. The second category of stimulus sentences had an intact script but violated a schema. We created schema violations by leaving the expected structure of events intact but replacing a noun denoting an acted-on object with an associatively unrelated noun. For example, in the sentence, "He started the car with the ignition key" the phrase "ignition key" was replaced by an associatively distant word, such as "hairbrush."

Script-breaking and schema-breaking. The third category of stimulus sentences contained both script and schema violations. The actions in these sentences neither formed a familiar sequence of events nor did they bear a familiar associative relationship to one another: "The supermarket owner painted her scarecrow magenta." These sentences were constructed by starting with an initial, completely comprehensible sentence and replacing both nouns and verbs in the abovementioned ways.

Script-compliant, schema-compliant: Oversufficiency items. Finally, we created a fourth group of items by leaving both scripts and schemas intact but varying the sufficiency of a particular action to fulfill the goal specified by that action. This group of items includes actions that "overshoot" their apparent motivation. For example, "The vacationers brought six hundred

cases of beer to the beach." Although these items should still seem puzzling to participants, they were designed to be less puzzling than the other three categories.

Pretest: Strangeness ratings. We created an initial pool of 40 puzzling behaviors (10 within each of the four categories), which were pretested on Mechanical Turk to verify perceived strangeness. Script-breaking items showed the highest overall strangeness ratings (M = 5.65), followed by items that were both schema- and script-breaking (M = 5.13) and items that were schema-breaking but script compliant (M = 5.03). The over-sufficiency category had the lowest overall strangeness ratings (M = 4.89).

We selected a subset of 16 items that (a) were representative of the means within item category, (b) provided a range of strangeness ratings, and (c) were low on another pretest dimension: laughability (stimuli that are so strange that they are seen as laughable, not cognitively puzzling).

Participants. Seventy participants (34 female) completed the study online through Amazon Mechanical Turk in exchange for monetary compensation. One participant was excluded for providing an invalid ID number. The average participant was 30 years old. Fifty-nine percent of participants had completed a two-year college degree or higher level of education.

Procedure. Participants were instructed that they would be reading a series of sentences. In response to each sentence, they were told to "add whatever sentences or phrases you think are needed to make sense of the sentence." Participants were instructed not to negate information in the original sentences but to "imagine a world in which the sentence is true." After adding their information they answered a follow-up question: "Taking into account the information you

added, how much sense does the situation described in the original sentence make now?" Participants rated their responses on a scale from 1 (No sense at all) to 8 (Perfect sense). We included satisfaction ratings to encourage participants to generate genuinely satisfying explanations for the task, but these ratings themselves could not serve as objective measures of explanation quality.

Out of the total of sixteen items, each participant responded to four, selected according to a full Latin square design that generated 16 forms. The distribution of the items' strangeness was roughly equal across forms. One item from the schema-breaking category had to be excluded from data treatment and analyses because people's action interpretations and explanations revealed that it was ambiguous with respect to the agent who performed the behavior in question**.**

Data treatment.

Content coding of explanations. Participants' efforts to make sense of the puzzling stimuli were coded using the Folk Explanations (F.Ex) coding scheme (Malle, 1998, 2004). Among other things, this scheme categorizes explanations into reason explanations and causal history of reason (CHR) explanations; and within CHR explanations, it distinguishes between stable traits (dispositional properties such as personality, character, or attitudes) and non-traits (such as roles, norms, and culture).

Reliability. Two independent coders classified free-response explanations to each of 209 puzzling behavior items, and a single coder classified the remaining 51 items. Coders reached 96% agreement on the codability of explanations ($\kappa = .74$). Coders reached 97% agreement on classifying explanations as reasons or causal histories ($\kappa = .74$), and 100% agreement on

classifying causal histories as citing either dispositions or other background information (κ = 1.0).

Validity. 12 participants failed to give a single valid response. Out of the remaining 57 participants' responses, 70% were valid for coding purposes. Invalid responses included those that negated the original sentence, merely repeated its content, or provided an outcome of the scenario described in the sentence instead of explaining it. To examine the role of strangeness in causing invalid responses, we conducted a logistic regression at the item level predicting validity scores from strangeness scores. Strangeness did aid prediction, with stranger items leading to fewer valid responses (Wald statistic = 13.41, $p < .001$, semi-partial $r = .21$).

Calculating explanation parameters. Using the F.Ex. system, participants' explanations were classified as either expressing a reason (a mental state) or a causal history factor (a non-mentalistic background cause) and, among the latter, a trait or nontrait. For each behavior that a person explained, we counted that person's number of mentioned reasons, causal history factors, and traits. For example, if, for a given behavior, a participant offered two reasons and one causal history (and it was a trait), that participant would have a score of 2 on the reason parameter, a 1 on the causal history parameter, and a 1 on the trait parameter.

2.3.2 Results and Discussion

Participants explained the puzzling behaviors in Study 1 on average with 0.89 reasons (SD = 0.40) per behavior, a rate that did not differ from the average of the six studies that represent our comparison standard, in which people explained ordinary behaviors (M = 0.98, SD = 0.64), Welch's $t(80.75) = 1.51$, $p = .13$, d = -0.14, 95% CI [-0.41, 0.13]. By contrast, participants offered 0.36 causal histories per behavior, which was lower than in the comparison studies (M = 0.71, SD = 0.75), Welch's $t(79.59) = 5.00$, $p < .001$, d = -0.47, 95% CI [-0.74,

-0.20]. To test hypothesis T, we analyzed people's rate of trait explanations within the causal history category (which makes the test orthogonal to that of causal history explanations), including both enduring personality traits and dispositional mental states (e.g., "He has always dreamed of returning to his childhood home"). Participants gave 0.08 traits (and 1.18 non-traits) per causal history explanation, which was considerably smaller than the average across previous studies in which participants explained ordinary behaviors (M = 0.50), $t(45.10) = 8.06$, $p < .001$, d = -0.93, 95% CI [-1.30, -0.56].

Study 1 provided initial evidence for Hypothesis M—that people routinely search for an agent's reasons (mental states) even when trying to make sense of genuinely puzzling behaviors. People's rate of reason explanations for these puzzling behaviors was as high as the rate of reasons explanations for ordinary behaviors; and instead of resorting to traits (or other background factors), people actually gave fewer of these explanations. The low rate of trait explanations is especially noteworthy in light of the familiar prediction that particularly unusual (expectancy-violating) behaviors should elicit an increase in dispositional inferences (Jones & Davis, 1965; Ross & Nisbett, 1991). However, our participant population consisted of Amazon Turk workers who may be particularly motivated or adept at solving puzzles. We therefore sought to replicate the finding from Study 1 in a sample of participants recruited from the Providence, RI community.

2.4 Study 2

2.4.1 Method

Participants. Participants were solicited via community advertisement and drawn from an existing database of non-student members of the Providence, RI community. A sample of 54

participants (29 women) completed the study online. The average participant was 38 years old, and 61% of participants had completed a 4-year college degree or higher level of education.

Procedure and Material. Participants received a URL for the study, whose procedure was identical to that of Study 1. The item that had to be removed from Study 1's analysis was replaced.

Data treatment.

Reliability. Two independent coders used the F.Ex. coding scheme to classify free-response explanations to each of 212 items, and an additional 4 items were classified by a single coder. Coders reached 97% agreement on classifying explanations as reasons or causal histories ($\kappa = .72$), and 100% agreement on classifying causal histories as citing either traits or non-trait background information ($\kappa = .77$). Agreement on the codability of explanations was 95% ($\kappa = .69$).[3]

Validity. Six participants were excluded because they failed to give a single valid response (e.g., they either negated or merely repeated the stimulus sentence, or provided outcome information rather than an explanation). Out of the remaining 48 participants' responses, 70% were valid for coding purposes. To examine the role of strangeness in causing invalid responses, we conducted a logistic regression at the item level predicting coding validity from strangeness scores. Stranger items did lead to fewer valid responses (Wald statistic = 6.72, $p = .01$, semi-partial $r = .14$).

Since the 16 puzzling behaviors were distributed across 4 forms, not every participant received every item. We therefore aggregated the score on each explanation parameter (reason, causal history, disposition) for a given behavior across all participants who had received that

[3] The substantially greater number of reasons over causal histories leads to a "penalty" in the size of kappa, which is highly sensitive to base rates.

item. Behavior strangeness correlated weakly with the number of trait explanations, $r = .17$, $p = .57$, *ns*.

2.3.2 Results

Participants offered 1.06 reasons ($SD = 0.41$) in response to the puzzling behaviors in Study 2, a rate that, as in Study 1, did not differ from comparison studies in which participants explained ordinary behaviors ($M = 0.98$, $SD = 0.64$), Welch's $t(62.88) = -1.30$, $p = .20$, *ns*, $d = 0.13$, 95% CI [-0.16, 0.42]. Study 2 participants also used fewer causal history explanations ($M = 0.39$, $SD = 0.37$) than did people in previous studies ($M = 0.71$, $SD = 0.75$), Welch's $t(75.15) = 5.29$, $p < .001$, $d = -0.44$, 95% CI [-0.73, -0.14]. Figure 2.1 summarizes the results from both Study 1 and 2, consistently demonstrating that people explain puzzling behaviors, compared with ordinary behaviors, with as many reasons but fewer causal histories.

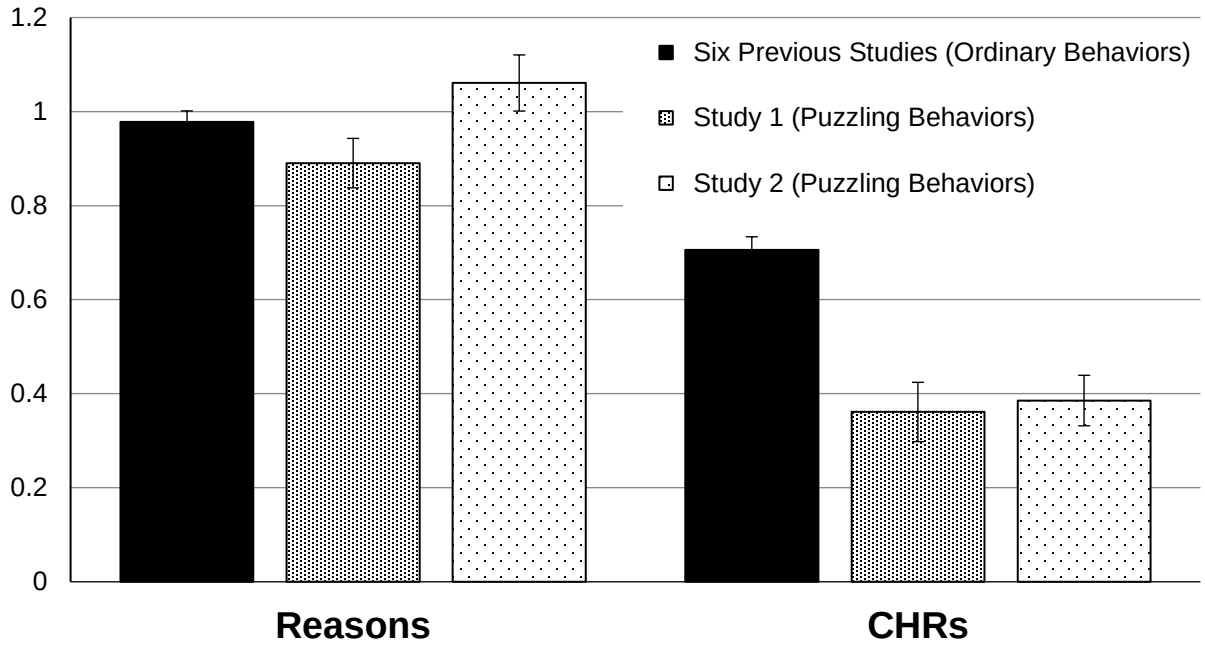

Figure 2.1. Number of reason explanations and causal history explanations (CHRs) for puzzling behaviors (in Studies 1 and 2) compared with ordinary behaviors (across six previous studies).

Traits. Once again we analyzed trait explanations (per behavior explained by causal histories). As shown in Figure 2.2, participants offered 0.16 traits (and 1.00 non-traits) when explaining the puzzling behaviors, rates that are considerably lower than those in previous studies, when people explained ordinary behaviors (M = 0.50, SD = 0.46), $t(38.73) = 4.98$, $p < .001$, d = -0.76, 95% CI [-1.12, -0.39].

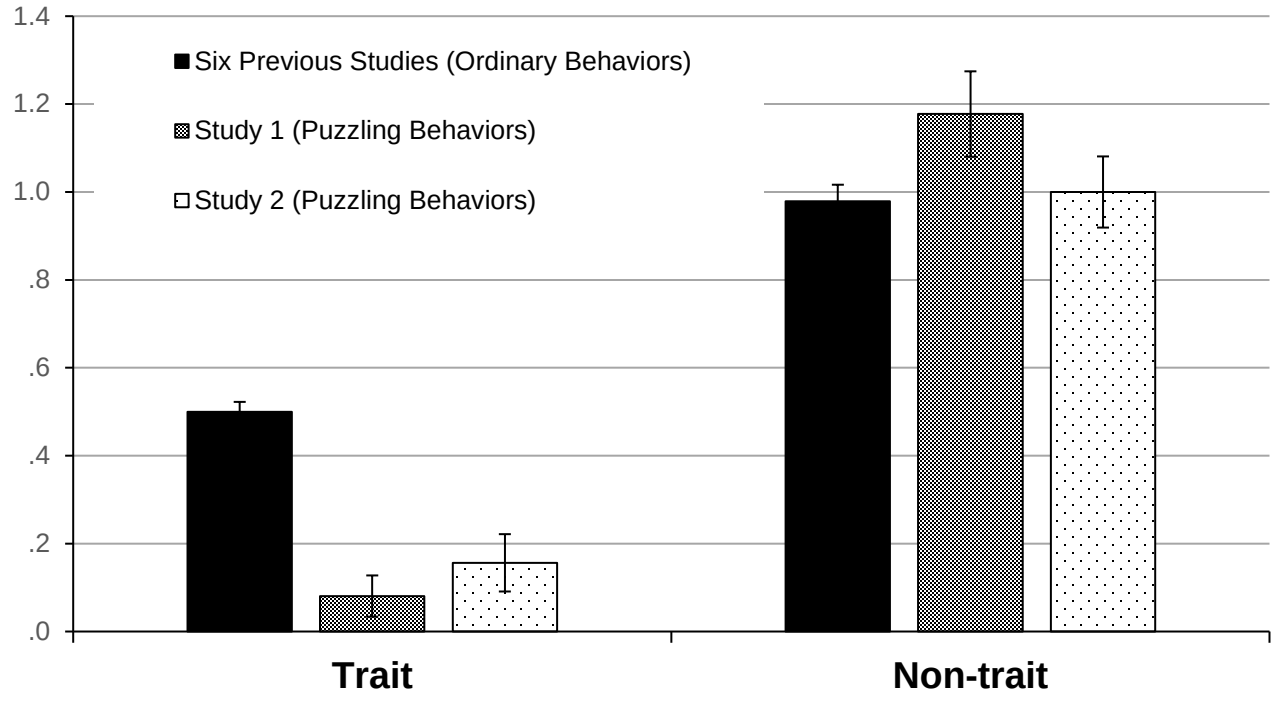

Figure 2.2. Number of trait (and non-trait) explanations given for puzzling behaviors (Studies 1 and 2) compared with ordinary behaviors (across six previous studies).

2.4.2 Discussion

Study 2 provided further evidence for Hypothesis M—that people routinely use mental state explanations even for puzzling behaviors—this time with a sample of community members. Participants persisted in providing the same number of reasons in response to puzzling behaviors as participants in previous studies did in response to ordinary behaviors. Furthermore, in both Studies 1 and 2, participants provided *fewer* causal history explanations and fewer trait explanations in response to puzzling behaviors than did previous participants when explaining ordinary behaviors. This result suggests that, when required, participants can and do go beyond their everyday scripts and schemas to actively search for and generate reason explanations, but

they are actually less willing or able to offer traits and other causal history explanations when faced with puzzling behaviors. Why might this be?

One possible account is that, lacking a clear knowledge structure for the puzzling stimuli, participants could no longer generate traits and other causal history explanations. This account of knowledge-based generalizations is supported by previous work on explanations of group behavior (O'Laughlin & Malle, 2002). When people explained the behavior of groups for which there is general, stereotyped knowledge available (e.g., members of a particular university's track team) they made ample use of causal history explanations. By contrast, when people explained the behavior of abstract groups about which they had no general knowledge (e.g., "Group A"), they decreased their use of causal history explanations and appealed to reasons instead. A lack of knowledge structures may also have posed specific problems for trait explanations in our first two studies. According to Jones and Davis (1965), inferences about people's dispositions are "correspondent inferences." That is, behaviors are seen as manifestations of more general traits that correspond in content or meaning to the specific behavior (e.g., "John helped the old lady cross the street; he did that because he's generous."). But because the strange behaviors in Studies 1 and 2 do not have a clear meaning, they are not easily understood as manifestations of any particular trait. Thus, participants may have struggled to generate traits.

2.5 Study 3

To address the concern that the strangeness of actions in our first studies negated common knowledge structures, we constructed new stimuli for Study 3, ones that are puzzles but nonetheless activate meaningful knowledge structures. Instead of puzzling *actions,* we presented participants with intelligible actions that were paired with puzzling *reasons* for those actions, a

stimulus that more closely approximates the type of contextualized puzzles people may encounter in real life. For example, people rarely encounter an action such as "The garbage men dropped off bags of trash at the end of each driveway" in isolation. Instead, they might first see the garbage truck passing by on its usual route and, within this familiar context, notice that the garbage man's behavior seems to pursue an unusual goal. A stimulus sentence such as "The garbage men drove their truck all around the city to drop off a can full of trash at the end of each driveway," provides both familiar context while also presenting a clear puzzle.

In addition, the presence of a puzzling reason explanation—rather than a puzzling action—may invite explainers to explain the reason itself by way of causal history factors, such as stable traits. This prediction is consistent with Jones and Davis (1965), who argue that traits are inferred directly from information about the actor's mental states underlying an action. Thus, if people frequently respond to puzzling behaviors with traits, but only when meaningful action or mental state information is available, Study 3 should capture this prerequisite.

2.5.1 Methods

Stimulus Development. Study 3 stimuli contained an initial clause that described an easily intelligible action (e.g., "The man went to the toy store. . . ") that was paired with a puzzling reason for that action (e.g., ". . . because he wanted to get some watermelons."). Stimulus reasons were either beliefs ("she thought that…," "she knew that…,") or desires ("He wanted to…," "so that he could…"). Thus, the puzzles in Study 3 stemmed not from any strangeness of the action itself (as in Study 1) but from the incompatibility between the action and its stated reason. Pretest ratings of such "action-reason incompatibility" provided a graded index of how puzzling each stimulus was. As in Study 1, we constructed four puzzle types,

violating either schemas, scripts, or both, or making the action (in light of its reason) seem oversufficient. Sixteen items of each type yielded a total of 64 items.

As in Study 1, we used a word replacement strategy to create items that fell into four "puzzle types": items that were script-breaking (e.g., "She put on a pair of flip flops because she wanted to improve her cattle-herding skills"), schema breaking (e.g., "She went to the plant nursery to pick up some video games for her son"), script- and schema- breaking (e.g., "She got her dress hemmed because the faucet was dripping") or items that violated neither a script nor a schema but indicated unusual means (e.g., "The man rented three semi trucks because he was transporting his dining room set to his new apartment").

In pretests, items that were both schema- and script-breaking showed the highest overall action-reason incompatibility (M = 5.63), followed by items that were schema breaking but script compliant (M = 5.24) and items that were script-breaking but schema-compliant (M = 4.82). As expected, the items with the lowest overall incompatibility were of the "oversufficiency" type (M = 4.05).

To convincingly place the puzzle in the action-reason pair rather than in the action on its own, we pretested a larger number of candidate pairs for the intelligibility of their action ("How well do you understand this action?") and eliminated ones that scored below 4 on a 1 (not at all) to 7 (very well) scale. Additionally, we eliminated pairs in which the intelligibility of the action alone was only barely higher than the intelligibility of the action-reason combination. A final pool of 64 stimulus sentences resulted, with 16 per puzzle type. (For a complete list of stimuli, see the Appendix.) Their action-reason compatibility ratings ("Given this explanation, how well do you understand this action?") ranged from 1 to 6.67 (M = 4.96) on a 1 to 7 scale. So that

higher scores would reflect *incompatibility*, the ratings were transformed by subtracting them from 8.

In the experimental task, each participant was presented with 2 practice trials and 32 stimulus sentences. From the total 64 items, eight distinct forms of 32 sentences were drawn, each including eight items from each of the four puzzle types, four with belief reasons and four with desire reasons. Items within each form had similar distributions of action-reason compatibility.

Participants. Forty-one members of the Providence, RI community participated in exchange for monetary compensation. They were recruited for study participation via direct telephone solicitation, bulletin board ads, and online advertising. Of the 40 participants who reported demographic information, 58% were female, and the mean age was 36 years. 60% of participants had completed a 4-year college degree, and 95% had completed at least some college. One participant who demonstrated insufficient comprehension of the experimental materials was eliminated from the analyses. A second participant whose accent was difficult to understand on the audio recording was also eliminated.

Procedure. Participants sat at a computer and wore a headset with microphone. The experimenter left the room and the participant read the instructions for the task on the screen. The experimenter then re-entered the room and confirmed the participant's understanding of the instructions. The participant then completed two practice trials and, after the experimenter left the room, began the main task.

In each trial, participants read a stimulus sentence at the top of the screen. After four seconds, a text prompt instructed them to "Add whatever sentences or phrases you think are needed to make sense of the sentence." Participants then had 45 seconds to provide their

additional information by speaking. As in Studies 1 and 2, participants were instructed to "imagine a world in which the sentence is true" and not to change or negate any details of the stimulus sentence but only to add information that helps make the sentence make better sense. After 30 seconds, they received a 15-second warning to finish up their spoken response for a given item. At the end of each trial they rated how well the original sentence now made sense in light of their added information (1 = No sense, 5 = Some sense, and 9 = Perfect sense).

Data Treatment

Response Validity. Out of 1248 total responses, 79% were valid for coding purposes. Invalid explanations negated the original sentence, merely repeated stimulus content, or provided outcomes rather than explanations. To examine the role of action-reason incompatibility in causing invalid responses, we conducted a logistic regression at the item level, predicting invalidity from action-reason compatibility scores. Action-reason compatibility did not predict invalid responses (Wald statistic = 1.16, $p = .28$).

Coding Reliability. Two coders were involved in the coding of responses to the 1248 items. Using the F.Ex coding scheme, both coders independently classified responses to the same 480 items. The remaining 768 items were classified independently by one of the coders and then reviewed by the other. The two coders reached 99% agreement on classifying explanations as reasons or CHRs ($\kappa = .84$), and 97% agreement on classifying causal histories as citing either traits or non-trait background information ($\kappa = .76$). Agreement on the codability of responses was 96% ($\kappa = .65$).

2.5.2 Results

To evaluate the rates of (spoken) explanations for puzzling behaviors, we again aggregated a set of previous studies to serve as a reliable comparison standard for explanations of ordinary behaviors, this time for spoken explanations (Malle et al., 2007)[4]. The number of reason explanations for puzzling behaviors in Study 3 (M = 1.59, SD = 0.64) was indistinguishable from the comparison standard of reason explanations for ordinary behaviors (M = 1.44, SD = 0.68), Welch's $t(57.46) = -1.32$, $p = .19$, *ns*, $d = 0.23$, 95% CI [-0.12, 0.57]. Likewise, the number of causal history explanations in Study 3 (M = 0.72, SD = 0.44) was indistinguishable from the comparison standard of causal history explanations for ordinary behaviors (M = 0.60, SD = 0.54), Welch's $t(65.15) = -1.52$, $p = .13$, *ns*, $d = 0.23$, 95% CI [-0.11, 0.58] (See Figure 3.)

Traits. Among causal histories, the number of trait explanations (M = 0.38, SD = 0.25, versus 1.01 non-traits) for the puzzling behaviors in Study 3 did not differ from the comparison standard of trait explanations for ordinary behaviors (M = 0.31, SD = 0.29), Welch's $t(78.61) = -1.61$, $p = .11$, $d = 0.28$, 95% CI [-0.09, 0.65] (See Figure 4.)

[4] Three of these studies were reported in detail in this article; two of them were included in its meta-analysis.

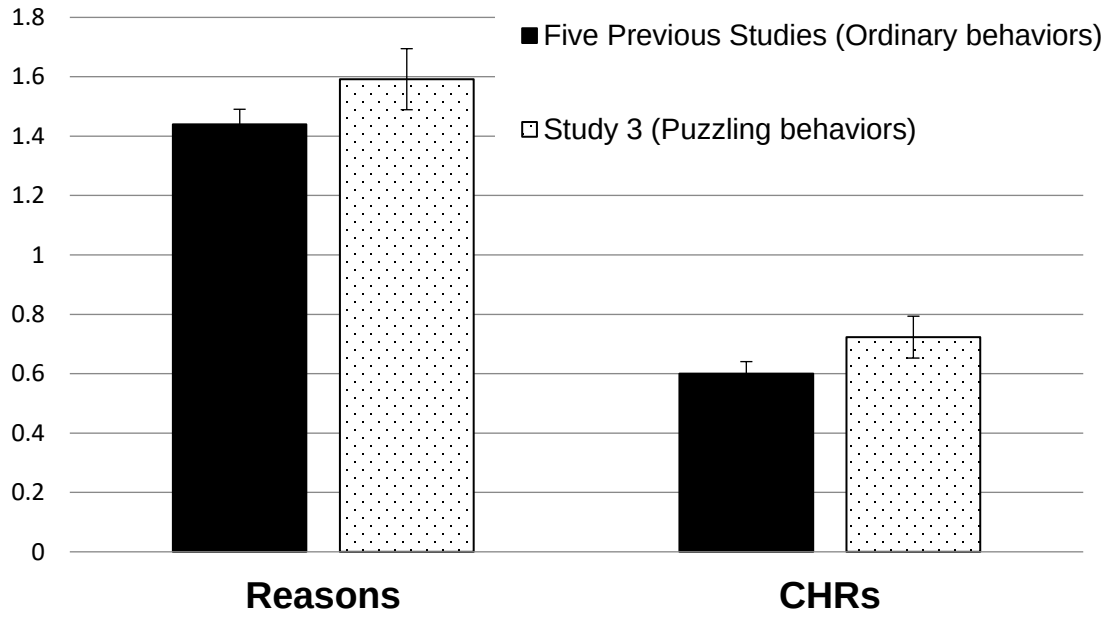

Figure 2.3. Number of reason explanations and causal history explanations (CHRs) given for puzzling behaviors (Study 3) compared with ordinary behaviors (across five previous studies).

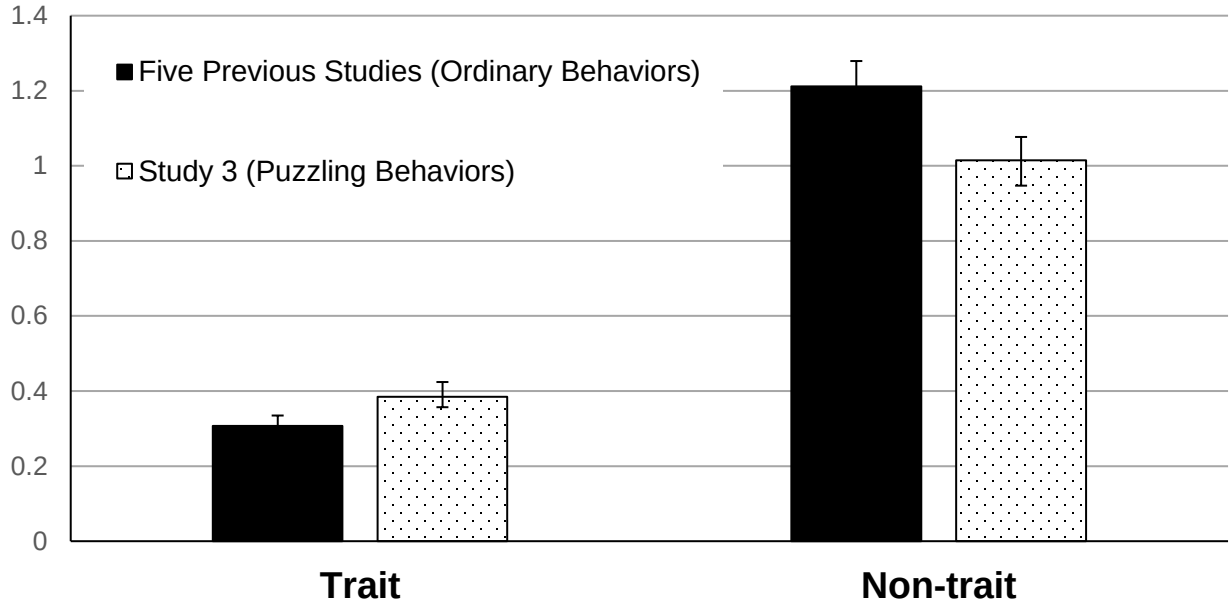

Figure 2.4. Number of trait (and non-trait) explanations given for puzzling behaviors (Study 3) compared with ordinary behaviors (across five previous studies).

Having available at least modest numbers of trait explanations, we examined whether the number of offered traits varied as a function of how puzzling the behaviors were (indexed by the action-reason incompatibility ratings). Since the 64 items were distributed across 8 forms of 32 items each, not every subject received every item. We therefore averaged the number of trait explanations for a given item across all subjects who had received that item. The more strongly incompatible (puzzling) an action-reason pair was, the fewer traits people offered for it (r = -.25, p = .05).

2.5.3 Discussion

As in Studies 1 and 2, participants in Study 3 persisted in providing reason explanations even for newly developed puzzling behaviors that contained additional information about the agent performing the behaviors. However, participants were able to generate relatively more causal histories than in the first two studies, supporting our conjecture that the single-sentence puzzling actions in Studies 1 and 2 had provided insufficient knowledge structures to afford causal history explanations. Even so, Study 3 participants offered no more causal history explanations than did the comparison samples when explaining ordinary behaviors.

Study 3 participants also increased their trait explanations relative to Studies 1 and 2, but once more, they offered no more traits than people normally do in response to ordinary behaviors. Moreover, the tendency to offer traits was negatively correlated with the degree to which a stimulus sentence was puzzling, thus further weakening the hypothesis that people select trait explanations for the purpose of clarifying puzzling behaviors.

2.6 Study 4

In light of Studies 1 through 3, evidence is mounting for the hypothesis that people routinely search for an agent's reasons (mental states) even when trying to make sense of genuinely puzzling behaviors. However, while Study 3 addressed some limitations of Studies 1 and 2, it came with one limitation of its own: the presence of a reason explanation in the stimulus sentence itself. We had designed this type of stimulus (locating the puzzle in the agent's mental state) in order to provide sufficient information for inferences about causal background and to perhaps even invite causal history explanations as accounting for the very reason that was puzzling. It is possible, however, that participants were effectively primed to provide more reason explanations because they encountered reasons already present in the stimuli. This account would not elucidate why reason explanations were frequent already in Studies 1 and 2 (neither of which featured reasons in the stimulus), nor would it elucidate why CHR explanations increased from Studies 1 and 2 to Study 3. But one might suspect that CHR explanations might further increase, and reason explanations decrease, if the stimulus behaviors contained incompatible *causal history of reason* explanations. To address this possibility, we presented participants in Study 4 with a new set of stimulus sentences that paired actions with puzzling causal history explanations.

2.6.1 Methods

Participants. One hundred participants (43 female) completed the study online through Amazon Mechanical Turk in exchange for monetary compensation. Participants' mean age was 35 years, and 60% of them had completed a two-year college degree or higher level of education.

Stimulus development. In naturally occurring explanations, causal history explanations can appear either by themselves or in tandem with reasons (Malle, 2004). We wanted to represent both variants in this experiment and therefore created two categories of stimulus sentences. In the first, an intelligible action was paired with a single causal history explanation (single-CHR items) that was incompatible with the action. For example, "The journalist shot photos of the crime scene [action] *because he always puts them on his nightstand* [causal history explanation.]" In the second category, an intelligible action was paired with both a reason that was incompatible with the action and a causal history explanation that functioned as a causal history of that particular reason (reason+CHR items). For example, "She went to the plant nursery *[because she wanted] to pick up some video games for her son* [incompatible reason]; *she was an indulgent parent* [causal history of that reason]."

To create maximally representative causal history explanations, we generated the reason+CHR items in two distinct ways. Half of the items were researcher-generated to capture the conceptual meaning of this explanation type (Malle, 1999). We designed the causal history explanations to provide a reasonable background for the stated reason, while still leaving intact the puzzle in the action-reason pair. For example, in the item above, being an indulgent parent provides a reasonable background explanation for why a mother would want to buy video games for her son, but it does not explain why she would want to get them at the plant nursery. The other half of the reason+CHR items were participant-generated—drawn directly from the subset of participant responses to Study 3 that included only causal history explanations. These were causal history explanations that specifically tried to solve the puzzles posed by incompatible action-reason pairs. For example, a Study 3 participant received the item, "The tax collector knocked on the family's front door because he wanted to get some chocolate," and gave the

response, “He knew the family [causal history explanation].” Our new, participant-generated item for Study 4 therefore read: “The tax collector knocked on the family’s front door because he wanted to get some chocolate; he knew the family.”

We created and pretested four single-CHR items and ten reason+CHR items (five researcher-generated, five participant-generated). Items in the single-CHR category had the highest degree of action-explanation incompatibility (M = 5.20), followed by researcher-generated reason+CHR items (M = 5.07). Participant-generated reason+CHR items had the lowest level of incompatibility (M = 4.07).

For the final stimulus pool we selected 4 items from the single-CHR category and 8 items from the reason+CHR category, including 4 that were researcher-generated (to reflect the theoretical meaning of this explanation type; Malle, 1999) and 4 that were participant-generated (drawn directly from explanations in Study 3). All reason+CHR items were formulated in two versions: mentioning the reason first or mentioning the causal history first.

Procedure. Instructions and procedure followed those of the previous studies. Each participant responded to one half of the item pool: two single-CHR items and four reason+CHR items (two generated by the researchers, two generated by previous participants), with order of item type counterbalanced across participants.

Data Treatment.

***Reliability*.** Agreement on the codability of explanations was 92% (κ = .52).[5] Two independent coders classified free-response explanations to 100 items. Coders reached 99% agreement on classifying explanations as reasons or causal histories (κ = .83), and 100% agreement on classifying causal history explanations as dispositions or not dispositions (κ = 1.0). The remaining 489 items were divided between the two coders.

[5] See 1.

Validity. Two participants failed to give a single valid response. Out of the remaining 577 responses given by 98 participants, 81% (469) were valid for coding purposes. To examine the role of action-explanation incompatibility in causing invalid responses, we conducted a logistic regression at the item level. The level of action-reason compatibility only marginally predicted the likelihood of invalid scores, Wald statistic = 2.92, $p = .09$, semi-partial $r = .04$.

2.6.2 Results

As shown in Figure 2.5, the number of reason explanations for puzzling behaviors in Study 4 ($M = 0.97$, $SD = 0.43$) did not differ from the number of reason explanations for ordinary behaviors in comparison studies ($M = 0.98$, $SD = 0.64$**)**, Welch's $t(161.01) = 0.15$, $p = .88$, *ns*, $d = -0.01$, 95% CI [-0.22, 0.20]. However, as in Studies 1 and 2, the number of causal history explanations for puzzling behaviors in Study 4 ($M = 0.46$, $SD = 0.37$) was lower than the number of causal history explanations for ordinary behaviors in the past ($M = 0.71$, $SD = 0.75$), Welch's $t(227.78) = 5.22$, $p < .001$, $d = -0.34$, 95% CI [-0.55, -0.12].

Puzzling behaviors with a single causal history explanation and puzzling behaviors with both a reason and a causal history explanation behaved exactly the same, showing lower causal history rates than ordinary behaviors elicited in past studies: for single-CHR items, Welch's $t(162.35) = 4.05$, $p < .001$, $d = -0.31$, and for reason+CHR items, Welch's $t(157.38) = 4.16$, $p < .001$, $d = -0.33$.

As in the previous studies, we aggregated the score on each explanation parameter (reason, causal history, trait) for a given behavior across all participants who had received that item. Action-explanation compatibility was not correlated with the number of trait explanations ($r = .024$, $p = .922$).

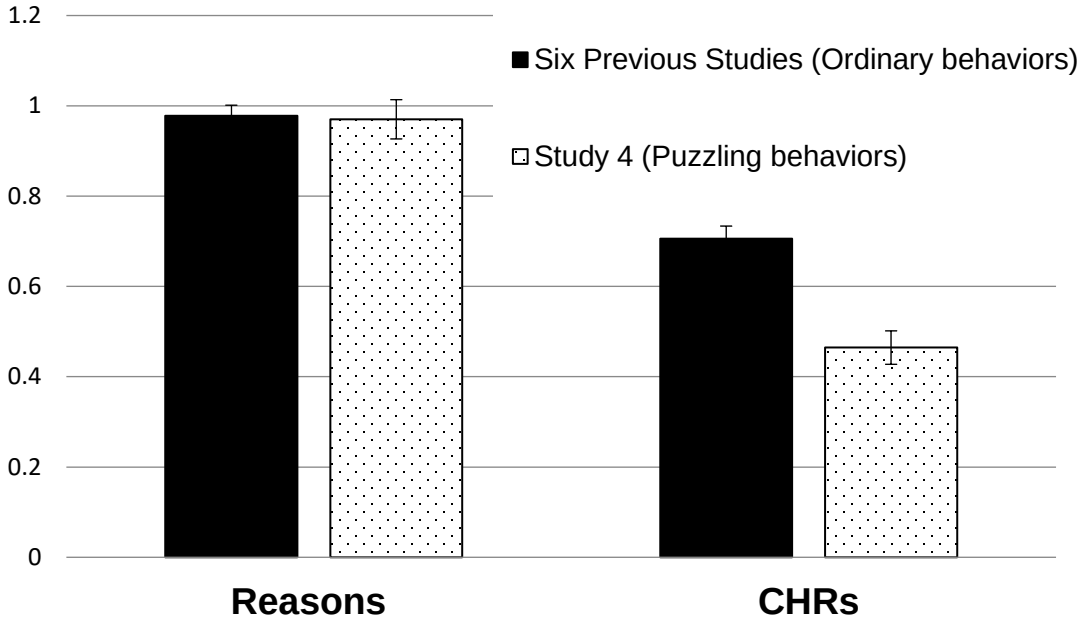

Figure 2.5. Number of reason explanations and causal history explanations (CHRs) for puzzling behaviors (Study 4) compared with ordinary behaviors (across six previous studies).

As shown in Figure 2.6, the number of trait explanations for puzzling behaviors in Study 4 (M = 0.15, SD = 0.29, versus 1.07 non-dispositions) was less than the corresponding number of trait explanations for ordinary behaviors in comparison studies (M = 0.50, SD = 0.46), $t(178.21) = 9.15$, $p < .001$, d = -0.81, 95% CI [-1.05, -0.57]. This difference held across both types of puzzling behaviors (single-CHR: d = -0.72; reason+CHR: d = -0.86).

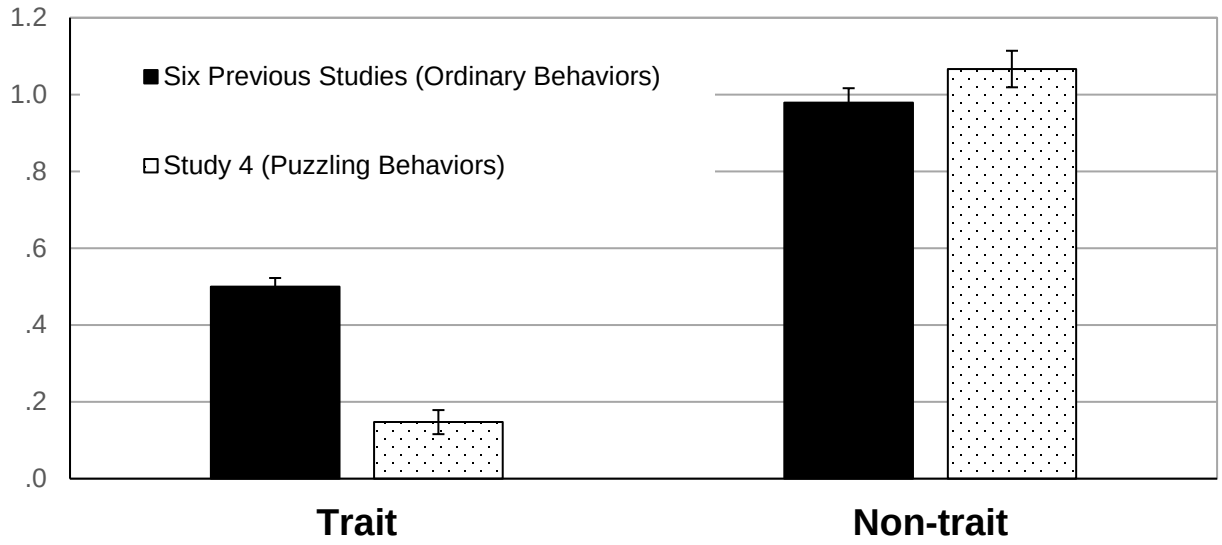

Figure 2.6. Number of trait (and non-trait) explanations for puzzling behaviors (Study 4) and ordinary behaviors (across six comparison studies).

2.6.3 Discussion

This study addressed the possibility that stimulus behaviors in Study 3 may have primed participants to provide reason explanations simply because each stimulus already presented a reason (that was incompatible with the behavior). We therefore constructed a new stimulus set in which causal histories (alone or in combination with a reason) were presented as part of the puzzling stimulus. But, confirming the patterns of all three previous studies, participants persisted in providing the same number of reasons (mental states) as people in the comparison sample did in response to ordinary behaviors. Moreover, despite being exposed to causal history explanations in the stimuli, participants did not offer any more such explanations in their responses.

In addition, even though the new stimuli were designed to provide information about the agent from which trait inferences could be constructed, the puzzling behaviors in Study 4 elicited

fewer trait explanations than did the ordinary behaviors in the comparison sample and fewer trait explanations than the puzzling behaviors in Study 3. This latter difference may derive from the difference between spoken and written explanations. When participants speak their explanations out loud as they did in Study 3, they provide more explanations overall (Malle et al., 2007) and, as a result, they also include somewhat more causal history explanations. In contrast, Study 4 elicited explanations in written form, and causal history explanations returned to their lower rate we had seen in Studies 1 and 2. The impact of communication mode (spoken vs. written) is just a small aspect of a larger phenomenon—the significant role that communicative forces play in shaping explanations (Hilton, 1990). For example, in the present studies the communicative audience was an unfamiliar experimenter, but systematic audience variations can lead to systematic variations in explanation types (Slugoski, Lalljee, Lamb, & Ginsburg, 1993). Whether there are any audiences that would substantially increase causal history explanations (perhaps sociologists or psychoanalysts) is a question for future research.

2.7 General Discussion

Across four studies, three unique stimulus sets, and online as well as local community samples, people overwhelmingly persisted in providing reason explanations in the face of puzzling actions, and, contrary to predictions of the classic trait attribution model (e.g., Jones & Davis 1965), people seem to offer no more traits or causal histories in response to puzzling actions than in response to ordinary actions. In fact, especially in the studies in which participants responded in the written medium, they struggled to generate trait explanations. This surprising pattern of results for trait explanations is illustrated in Figure 7, which places the

means for traits from the present four studies in the context of averages from past studies that made up our comparison standards.

2.7.1 Meta-analysis

To examine the reliability of our findings, we derived effect sizes (Cohen's *d*) for comparing the puzzling behaviors in each of the present studies and ordinary behaviors in previous studies for reasons, causal histories, and traits. In random-effects models using inverse variance weights, the average effect size of reasons was $d = .03$, 95% CI [-0.10, 0.16], $z = .40$, $p = .69$. No inhomogeneity was detected, Q (df = 3) = 3.44, $p = .33$. Thus people gave as many reasons in response to puzzling behaviors as they did in response to ordinary behaviors in previous studies, and the effect size approached 0. For causal histories, the average effect size across four studies was $d = -0.27$, 95% CI [-0.57, 0.028], $z = -1.18$, $p = .08$. However, heterogeneity was detected, Q(df = 3), $p < .01$. When response medium (written or spoken) was added as a moderator, $\gamma = -0.63$, $z = -3.35$, $p = .001$, Q(df = 2) = .68, $p = .71$, the average effect size for the three written studies alone was $d = -0.40$, 95% CI [-0.53, -0.26], $z = -5.62$, $p < .001$, yielding an overall estimate of a small to medium effect of people providing fewer written causal histories for puzzling than ordinary behaviors. For traits, the average effect size across all four studies was $d = -0.56$, CI [-1.10, -0.02], $z = -2.04$, $p < .05$. Heterogeneity found in this initial analysis was also accounted for by response medium, $\gamma = -1.10$, $z = -4.97$, $p < .001$, Q(df = 2) = .43, $p = .81$, and the average effect size for the three written studies alone was $d = -0.82$, $z = -.86$, 95% CI [-1.01, -.63], $p < .001$, yielding an overall estimate of a large effect of people providing fewer written traits for puzzling than ordinary behaviors.

A reanalysis of previous results (Malle, Knobe, & Nelson, 2007) showed that responding in a spoken medium had higher overall explanation rates, which helps explain the increased rates

of causal histories and traits in Study 3, relative to the written explanations in Studies 1, 2, and 4. The written format may thus present a higher threshold for explanatory activity or reporting. In addition, Study 3 differed from the three written studies in that community members took part in a laboratory. These participants may have been more motivated to produce explanations than community members who completed the study online or Mechanical Turk online workers. It is thus possible that if we replicated Studies 1, 2, and 4 in the lab and in the spoken medium, we might obtain results closer to that of Study 3, which showed no significant decrease in causal histories or traits, as the other studies showed.

Most important, the relative rate of reason explanations was robustly constant across all four studies, across both written and spoken, lab and online studies. This persistence of reasons (and decrease in causal histories and traits in all but one study) suggests that reasons are indeed the "default" explanatory tool—used whenever and wherever explanations are needed.

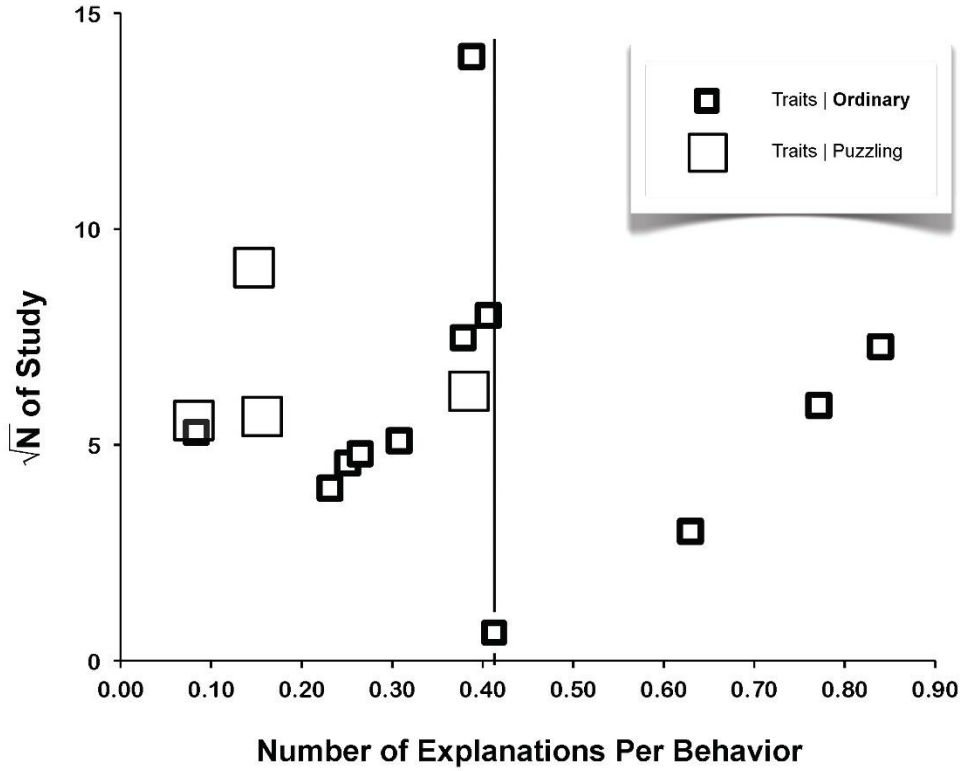

Figure 2.7. Entries show average number of trait explanations for puzzling behaviors in present four studies and for ordinary behaviors in comparison studies, including 5 spoken and 6 written explanation sets. Following practices in meta-analysis, the y axis indicates a weight for each mean (square root of the sample size on which the mean is based). The aggregated comparison standard (sample-size weighted mean across previous samples) is marked by a vertical line.

The consistent use of reason explanations even in the face of seriously puzzling behaviors supports Hypothesis M—which claimed that reasons (mental states) are the routine and default explanation mode for any intentional behavior, ordinary or puzzling. By contrast, the results do not support Hypothesis T—the claim that people would explain genuinely puzzling behaviors increasingly with traits (and other causal background factors). It also calls into question its corollary—that people would abandon explanations using mental states when familiar scripts and schemas (as provided by ordinary behavior) are no longer available. In fact, the lower prevalence of trait and causal history explanations in response to puzzling behaviors suggests

that it may be these explanations that rely heavily on familiar, general knowledge and are more difficult to generate when scripts and schemas are violated. By contrast, people had no difficulty constructing reason explanations under these informationally sparse conditions.

2.5.2 Frequency and Function of Traits in Behavior Explanations

The infrequent use of dispositional explanations in response to puzzling behaviors suggests that dispositions may be used less commonly than previously thought (here, they made up 7.5% of all explanations across the four studies). In particular, their low prevalence in the face of puzzling behaviors contradicts the common hypothesis that when behaviors deviate from expectations or normality, social perceivers are particularly inclined to make trait inferences (Gilbert & Malone, 1995; Jones & Davis, 1965; Ross & Nisbett, 1991). The literature considers this hypothesis to be very well established (Fiske & Taylor, 1991); how can we reconcile our finding with the previous literature?

We offer one theoretical and one methodological point of reconciliation. The theoretical point is that trait *explanations* are psychologically distinct from trait *ascriptions*. The latter are likely to occur when the perceiver is motivated to form an impression of a target person (Hamilton, 1998) and especially when such an impression is to be communicated to others. Impressions are particularly useful as summary information, and people are prone to use them for broad predictions (e.g., of job performance over an extended time). But such general person impressions, and the trait inferences that constitute them, are not as useful in the domain of action explanation, when a social perceiver is trying to make sense of a particular behavior that someone performed. Making sense of puzzling behaviors is just what people in our studies were asked to do and, as in previous research on ordinary behavior explanations, they explained those puzzling behaviors predominantly with mental states. It is quite possible, of course, that a study

with similar stimuli but a task that emphasizes understanding the whole *person*—for example, a processing goal to form an impression (Hamilton, 1998) or anticipate a future interaction (Devine, Sedikides, & Fuhrman, 1989)—could produce a different result: a greater reliance on traits. However, there was nothing in the present task or instructions that *prevented* people from mentioning traits. The rarity with which they did so suggests that behavior explanation and trait ascription are two distinct processes.

The methodological point is that people's trait ascriptions may have been augmented in past studies by selective stimuli (specifically designed to elicit trait inferences; Uleman et al., 2008) and selective response options (rating scales of dispositions only; e.g., Jones & Harris, 1967). Under these circumstances, studies document high levels of dispositional attributions, whereas more naturalistic studies (without tailored stimuli or limited response options) find surprisingly low levels of dispositional attributions (Lewis, 1995). The present studies did not rely on stimuli tailored to elicit particular explanations but employed a full range of stimuli: actions alone, actions with reasons, actions with causal histories, and actions with both reasons and causal histories. Moreover, the present studies left response options unconstrained: people offered explanations in their own words (written or spoken), with no instructions except to make sense of the behaviors. Under such conditions, we learned, people tend not to spontaneously provide traits but readily offer mental states.

2.7.1 The Unique Function of Reason Explanations

What can reason explanations accomplish that trait explanations (and causal history explanations more generally) cannot? Explaining another person's behavior by appealing to a causal history explanation such as a trait almost always explains an action by subsuming it under a broader pattern that holds for a type of agent or for a type of action. In contrast, reasons are

individuated along three dimensions: They are specific to a particular agent (the agent who possessed the mental states), a particular context (the reasons were on the agent's mind at that time and in that situation), and a particular action (they serve as the rational grounds for that particular action). Using a causal history explanation we might, for example, explain one man's saving a cat from a tree by saying that "he is a firefighter." This serves as an explanation in virtue of the knowledge that *saving cats from trees is something firefighters tend to do.* But as we have seen, when such generalizations are absent or contradictory, people search primarily for mental states to make sense of others' behaviors. In contrast to causal histories, reasons provide information about the particular person performing the particular behavior in the particular context. Rather than answering questions like, "Why might a person save an animal?" reason explanations answer questions like: "Why did *this person* save *this cat* from the tree *today?"* Perhaps he knew that the cat belonged to his grandmother's friend (reason), and he figured that since it was a slow day at the firehouse, he could use his equipment to help (reason).

While causal histories derive their explanatory power from invoking familiar generalizations, reasons are more flexible, enabling novel combinations of knowledge about firefighters, grandmothers, a slow day at work, and even the less plausible combinations of social scripts and roles found in our stimuli. For example, in response to the stimulus sentence, "She put on her flip flops because she wanted to improve her cattle-herding skills," one response was: "She figured a pair of comfortable shoes would make her more relaxed and possibly cause injury. [reason] By using flip flops she had to be more aware of her surroundings and take each step carefully. [reason]" Knowledge-based generalizations are not available as tools for the explainer in such a unique, novel case. And neither is social projection—another common tool to understand other people's behavior (Clement & Krueger, 2000)—that requires some

meaningful basis for similarity between the explainer and the agent. The only readily available tool in such a case is a *simulation* of the agent's mind (Nichols & Stich, 2003). In a simulation, the explainer considers the situation in which he himself might *wear flip flops to go cattle herding* and how flip flops might possibly improve his cattle-herding skills if he did. It is only through special consideration of this individuated counterfactual scenario that the explainer is able to find a novel connection between uncomfortable shoes and a challenging cattle-herding experience. Because simulations can easily individuate information along all three dimensions discussed earlier (agent, context, and action), they are the perfect tool to construct reason explanations, especially for puzzling behaviors.

We should caution, however, that our studies have highlighted the types and frequency of explanations people offer in response to puzzling behaviors; they have not addressed the quality of these explanations. In two related studies (Korman & Malle, 2016) we have explored this issue by turning the present participants' explanations into experimental stimuli and presenting them to a new group of lay perceivers. In keeping with the finding that reasons are of fundamental importance, these independent perceivers judged reason explanations as enhancing understanding better than causal history explanations.

2.7.2 Action Explanations and Other Kinds of Explanations

The predominance of particularized explanations for novel, puzzling instances appears to be unique to lay explanations of human action. Both philosophical treatments of scientific explanation (Hempel, 1966; Kitcher, 1989) and empirical evidence from children and adults' explanations of nonbehavioral events (Lombrozo, 2009; Walker, Lombrozo, Legare, & Gopnik, 2014) suggest that explanations that appeal to a general pattern (Williams & Lombrozo, 2010) are the most highly favored. This may be in part due to their function of facilitating predictions

about similar phenomena in the future. In everyday life, however, a person's most immediate concern about another person's puzzling behavior is not forming a predictive generalization (the domain of causal history explanations) but understanding a particular present behavior (the domain of reason explanations) and planning the proper response to it.

2.7.3 Simulation, Reasons, and Their Limits

The simulation-based reason-giving our participants engaged in is not an inevitable response to everyday behaviors. As previous research suggests (Epley, Keysar, Van Boven, & Gilovich, 2004; Lin, Keysar, & Epley, 2010), people do not engage in simulation and reason-giving all the time; this would be cognitively expensive and often plain unnecessary (in many cases scripts and norms do a very fine job). But in response to many significant puzzles they face in their everyday lives, people are able to go beyond their knowledge base and invest in simulation-based mental state inference. So future work is needed to examine exactly when this more cognitively expensive processing is turned on and off and what immediate payoffs it has in dynamic social interaction.

Moreover, some puzzles may activate the simulation machinery without producing reasons, such as excessively altruistic self-sacrifice, suicide, or extreme violence. People struggle to understand, for example, why a family might go so far as to adopt 20 children: "Some people thought they were saints; but others thought they were publicity-seekers, or weirdos, or had some kind of psychological disorder. Some thought they were addicted to acquiring kids to fill some need, the way others were addicted to shopping" (MacFarquhar, 2015). All of these are causal history explanations—vague generalizations borne of the inability to simulate the agents' actual reasons, as even people who "thought they were saints couldn't understand why they did it" (MacFarquhar, 2015). Likewise, commenting on the perpetrator of the Sandy Hook

massacre, Solomon (2014) wrote that even if we discovered that Adam Lanza had suffered from schizophrenia or pedophilia, or had been abused as a child, people "still wouldn't know why he acted as he did"—that is, they wouldn't know his reasons. Nonetheless, finding themselves at the limits of their own explanatory capacity, people reveal their unbending tendency to grasp for *reasons*, even when none are in reach.

CHAPTER 3.

Mentalistic Rationality: The Differential Function of Belief and Desire in Intentional Action Explanation

3.1 Rationality in Action: Teleological vs. Intentional Stances

Infants and adults alike understand intentional actions by inferring the *goals* of those actions. They infer these goals by appeal to what researchers have dubbed the "teleological stance": an interpretive stance that relates the action itself to the goal object or location and constraints of the physical environment (Gergely & Csibra; 2003; Gergely et al., 1995). For example, when observing an agent traveling through an open environment without barriers toward a goal, both infants (after goal habituation) and adults expect the agent to take the most efficient path to the goal location. In fact, when presented with an action that moves inefficiently through the environment with respect to one physical goal, adults may infer that the goal has more likely changed to a different location – one for which that path is more efficient (Baker, Saxe, & Tennenbaum, 2009).

These same researchers (Gergely et al. 1995; Baker et al. 2009) have also argued that this teleological stance is analogous to another, richer stance held by adults: an *intentional* stance informed by a *rationality assumption* (e.g., Dennett, 1987; Davidson, 1963). Under this assumption, actions are performed to fulfill the agent's desires in light of the beliefs she holds. Like the teleological stance, the intentional stance relates three components to one another: actions, goals and constraints on the fulfillment of those goals. However, there are at least two key distinctions between the two stances.

First, unlike the teleological stance, the intentional stance is fully mentalistic. It refers not only to goals that can be easily "read" off of common actions in the physical world, but also to abstract, higher-order goals that must be inferred from those actions (i.e., *desires* – e.g., to feel happy, save the world, or trick someone). Indeed, for purposes of achieving the most complete comprehension of an action, higher-level descriptions of the desires fulfilled by an action are often preferred (Vallacher & Wegner, 1987). And an agent's beliefs refer not just to constraints on the physical environment but to any subjective belief about constraining circumstances on how the agent's action may fulfill her desires. Regardless of how an action may appear on the surface, a wide diversity of distinct beliefs may underlie it – including false ones.

The second distinction between the two stances is a consequence of the introduction of genuine mental state representations. In teleological reasoning, the goal itself (even though it takes place after the action, temporally) serves as the Aristotelian "final cause" of the action; that is, the agent is said to pursue a particular action along a particular path *so that* she can achieve a particular goal (Lombrozo & Carey, 2006). Thus, in the teleological understanding of action, the most important inference to be reached is clearly a goal inference (see measures used in Gergely & Csibra 1995; Baker et al. 2009; Schachner & Carey, 2013). In contrast, the full blown, mentalistic "rationality assumption" appears to rely on the unique informativeness of two distinct mental state representations for understanding: beliefs and desires. According to this assumption, the agent chooses to perform a particular action *both* because she has a *desire* for that outcome, and a *belief* that her action will bring about that outcome. Distinct from the 'final cause' status of goals in the teleological stance, the two mental states must be (or have been) on the agent's mind as *prior* causes of the action. And in addition to serving as causes of the action,

these mental states also subjectively rationalize the action in the mind of that agent, serving not only as objective causes of the action, but as subjective *grounds* for the action.

A long tradition of research on "theory of mind" investigated use of these mental state concepts in a variety of contexts. For example, research on the belief concept demonstrates people's ability to represent others' epistemic states, primarily through the study of beliefs with false contents or contents that differ from an agent's own current knowledge state (e.g., Hogrefe, Wimmer, & Perner, 1986; Wellman, Cross, & Watson, 2001; Wimmer & Perner, 1983). In contrast, desires are described primarily in terms of their functions in representing different goals for different individuals, even with respect to the same object or situation (e.g., Astington & Gopnik, 1991; Repacholi & Gopnik, 1997, Wellman & Wooley 1990,). Beliefs and desires are also distinct in their developmental trajectories (with desire coming before belief in development, e.g., Bartsch & Wellman, 1995) and in the degree to which they are grasped by primates (who come to grasp desire but never belief, Premack, 2010; Premack & Woodruff, 1978). However, many of these investigations have focused on cognitive functions served by these mental states other than action explanation, such as straightforward communication of one's own or another person's epistemic state or object of desire (Bartsch & Wellman, 1995). There is also evidence for a distinction in the *social* function of belief and desire in action explanation; people use the two mental states differently for purposes of impression formation, for example, explaining their actions with more beliefs if they seek to appear rational to others (Malle et al., 2000), or providing more desires if they seek to excuse a negative behavior (Korman, Cusimano, Smith, Monroe, & Malle, 2014). However, if beliefs and desires serve distinct cognitive functions in the explanation of intentional action in mature adult observers, these have yet to be demonstrated.

3.2 Belief and Desire: One and the Same?

While not exploring the differential functions of belief and desire directly, contemporary work on people's rational understanding of action does suggest one possible account. In work on the teleological stance, researchers argue that if the social perceiver can identify that any two of the three main components of action, goal, and environmental constraints are present, then the third should be inferable from these two (Gergely & Csibra, 2003; Baker at al., 2009). If beliefs and desires can be thought of as analogous to constraints and goals, respectively, then we might expect that a similar inferential rule would hold for belief, desire, and action. However, we found no evidence for this hypothesis in a previous study (Chapter 2, Study 3 of this dissertation). When we provided participants with intentional actions paired with either a belief or a desire, people offered the same number of beliefs and desires in their explanations regardless of which mental state was paired with the action in the stimulus, $F(1, 38) = 1.95$, *ns*.

One interpretation of this result is that, regardless of the differences in their conceptual structure as representations, beliefs and desires do not serve fundamentally different functions as explanations for human action. Indeed, previous work shows that people tend to give about one reason (belief or desire) explanation for every behavior they explain (*M* = .98 for ordinary behaviors, *M* = 0.97 for puzzling behaviors). This finding supports the notion that perhaps *both* mental states are not required, as a single mental state –*either* a belief or a desire – will often suffice to make the action intelligible to the explainer. This idea is especially plausible when we consider the fact that, in some cases, belief and desire may deliver the same types of contents in only very slightly different forms (Ross, 1977; Malle et al., 2000). Consider the following two

action descriptions: He gave chicken soup to his sick grandmother because he **wanted** her to feel better or …because he **knew** it would make her feel better. I will revisit this example below after introducing an alternative account of the functions of belief and desire.

3.3 Explaining Action with Mental States: Connecting Action, Means, and Goal

In the present study, we aim to explore an alternative hypothesis about the psychological determinants of explainers' selection of belief versus desire explanations for intentional action within a fully mentalistic framework. Even if, as previous work suggests, people only tend to draw on only one or the other explanation type for a single behavior, each type still may serve a distinct function in explanations of intentional action. Our theory relates three key elements of a fully mentalistic rationality assumption: every intentional action is explained by connecting the *action*, the agent's subjectively represented end *goal*, and the agent's representation of the action as a *means to the goal.* Following Dretske (1988) we argue that beliefs and desires draw on these elements differently. First, desires provide the crucial "destination" of the action – the end toward which the action is ultimately aimed (Dretske 1988, 132). Most fundamentally, a desire represents the goal that the agent has on her mind before acting, and that motivates her decision to act. Consistent with findings on the teleological stance in young infants, which focus on the identification of goals as primary in action understanding, desires also appear to be primary in adults' understanding of intentional action, and are inferred first, before beliefs, when adult social perceivers observe such actions (Malle & Holbrook, 2012; Haigh & Bonnefon, 2015).

In the presence of a clear goal, inferring the agent's belief provides the social perceiver with a subjective understanding of how the agent sees the action as a *means* to that goal, shedding light on an agent's decision to achieve her goal by performing that particular action. In

other words, beliefs provide the social perceiver with the "map by means of which an [agent] [has] steer[ed]" toward that goal (Dretske 1988, p. 79). For example, one participant provided this belief explanation to show how a man's action of *purchasing an expensive oriental rug* would serve as a means to the goal of *improving his vision*: "He thought that staring on the patterns on the rug would improve his vision." (Chapter 2, Study 4).

Let us return now, briefly, to the "chicken soup" example above, which suggested that belief and desires may serve identical conceptual functions after all. Such a simple action (giving chicken soup to one's grandmother) relies on widely accepted knowledge structures (social scripts) that link particular "destinations" (end goals) with particular "maps" (actions-as-means) toward those goals. The belief that connects the action to the goal (knowing that chicken soup will make Grandma feel better) already implies the destination (to make her feel better), and vice versa. In other words, the existing action description already delivers full comprehension (See Figure 3.1), so stating a belief or desire reason does not introduce any novel explanatory information. Therefore, if beliefs and desires have distinct conceptual features, these features will not be easily revealed in explanations for such simple actions. In contrast, when puzzling incongruities between agent's observed action, presumed goal, and choice of means are presented, these distinctions may be more likely to reveal themselves. Specifically, the present chapter investigates how beliefs and desires may serve as distinct answers to two types of explanatory puzzles (Bromberger, 1970; Hilton, 1990). These puzzles and the role of each explanation type in providing solutions are outlined in the following pages.

Figure 3.1. When an action and a belief or desire explanation with which it is paired already conform to a common knowledge structure, knowledge about the action's goal comes "for free" with knowledge of the means, and vice versa. To the social perceiver, the goal is clear, and the choice of means is appropriate. The action is thus comprehensible without addition of the missing "third element" (belief or desire). In this and all subsequent figures, a straight line to the goal in question is meant to represent the social perceiver's complete subjective understanding of the relationship between the action-as-means and the goal. In subsequent diagrams, breaks in the social perceiver's understanding are represented by jagged, indirect lines (means) from action to goal, or by lines that point toward the wrong goal.

3.3.1 Eliciting Belief Explanations: Means-End Puzzles

If beliefs are map-like indicators of means-end relationships, they should be especially apt answers to explanatory puzzles that arise when an action is difficult to understand as a **means** to its goal. In such cases, the goal is either explicitly stated or strongly (and uncontroversially) implied, and it is action's proposed *path to fulfillment* of the goal that is puzzling – not the goal itself. Notably, beyond its representation as a *physical* path to achievement that is perceived as either "efficient" or "inefficient" in reaching objects in space, the efficiency with which an action achieves its goal is determined by more general knowledge about the means people commonly choose for the achievement of (physical or nonphysical) goals. These **means-end** puzzles thus violate the typical means for the given goal (See Figure 3.2a).

For example, an agent who straps on her cross-country skis with the goal of making it to the supermarket on time is fulfilling this goal in an atypical fashion. We predict that in response to such puzzles participants will provide *belief* explanations, which will clarify how, in the agent's mind, the action actually serves as a means to the given goal. In response to this example, the social perceiver might respond that the agent *knew that the roads were blocked with*

snow and *she realized that the only way to get around was on cross-country skis*. This prediction is also consistent with findings in the action identification literature (Vallacher & Wegner, 1987), which suggest that people tend to focus on the details of an action (*how* it achieves its goal) when it pursues its goal by complex or unfamiliar means.

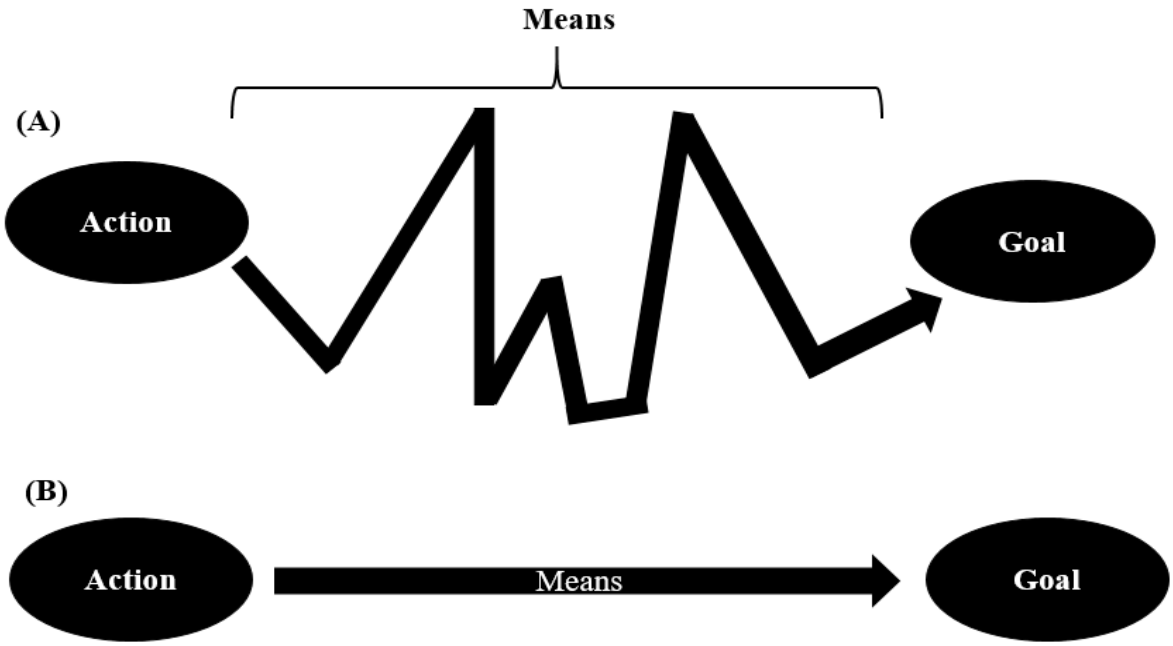

Figure 3.2. When the agent pursues her goal by a puzzling means as in (a), the social perceiver understands what the goal is, but not why the actor would choose such an atypical way to reach the goal. Although an alternative approach to the puzzle would be to posit many subgoals (additional desires) en route to the final goal, this approach is neither parsimonious (it likely requires multiple additional goals) nor illuminating: it still does not explain why the agent would have chosen a convoluted, multi-goal route to the original goal in the first place. Only a belief provides a clear solution: it "straightens out" the perceived jaggedness of line in (a) in the mind of the social perceiver, as in (b), by explaining how the agent's choice of means *is* an intelligible way to reach the agent's goal, *given the constraints she had in mind*.

3.3.2 Eliciting Desire Explanations: Goal-blocking puzzles

As discussed earlier, desires are the agent's mental representations of her goal. In many cases, the desire behind a person's action can be easily generated just by observing or hearing about that action; for example, people typically go to the pool because they want *to swim*, and to the theater because they want to *see a show*. However, if an action lacks a clear, single end goal,

people may need to provide desire explanations in order to clarify the action's final destination. Our ***Goal-blocking*** puzzles sought to create such ambiguity by presenting actions that contained multiple, contradictory goals.

Many actions can be described as having multiple goals. However, in order for these goals to all be goals of that *same* action, they must be arranged in a hierarchical, consistent structure such that there is only one end goal—the "destination" of the action, which is specified by a desire motivating the action at the highest level. This desire provides a structure under which each goal of the action is consistent with the other goals also achieved by that action. Consider the following: *Jane went to the department store. Her goal was to get her mother a present. She did that in order to make her mother feel happy on her birthday.* Jane's action fulfills a series of hierarchically related goals: getting the present is both a fulfilled goal and an action that fulfills further goals, with one *end* goal: to make her mother feel happy on her birthday. When an action is performed in service of multiple, *inconsistent* goals – when the pursuit of one goal actually runs *counter* to the achievement of another goal, thus effectively blocking its achievement in that same action – an explanatory gap opens up (See Figure 3.3). For example: "He hung up on his sister because he wanted her to know how much he loved her." Generally, one person hangs up on another person to show displeasure, and showing displeasure tends to be mutually exclusive with the demonstration of love for another person; that is, they cannot normally be executed in the very same action.

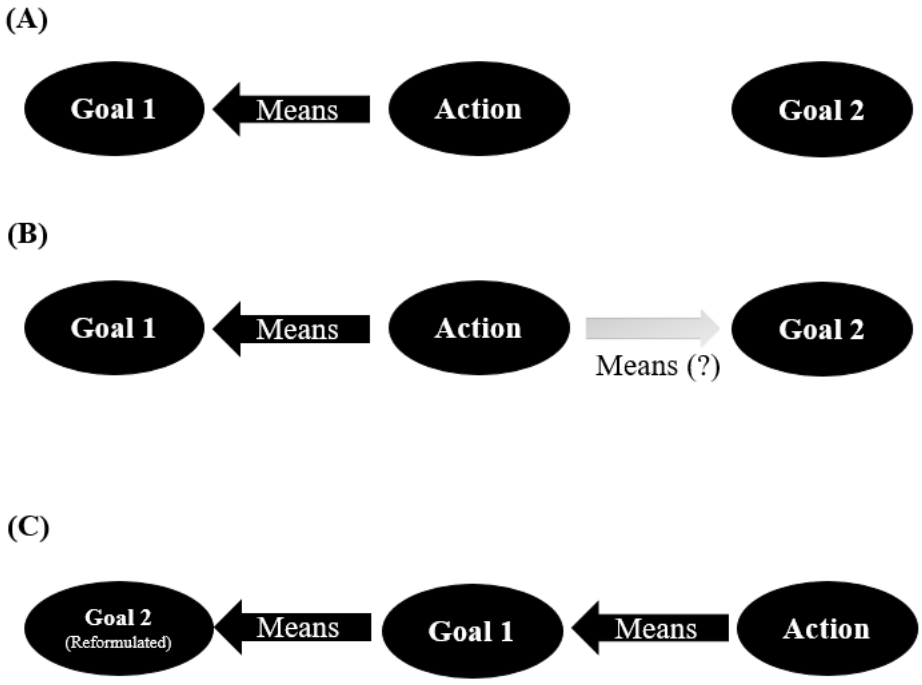

Figure 3.3.
(a) Schematic of a goal-blocking puzzle. An action that implies the fulfillment of one goal is presented, but the agent also wants that action to fulfill a second, contradictory goal.

(b) Belief "solution" to a goal-blocking puzzle. A belief may posit (rather uninformatively) that the agent falsely believes that the action serves as a means to the fulfillment of Goal 2, but this is unlikely to enhance the intelligibility of the action. It is still unclear how the *same* action that strongly implies Goal 1 could also achieve a contradictory goal. With an added belief connecting the action to Goal 2, the action now moves in two 'opposite directions.'

(c) Desire solution to a goal-blocking puzzle. A common way to resolve this puzzle would be to reformulate Goal 2 (without negating it) so that the action can achieve both goals by pointing in the same "direction". If the status of Goal 1 as a means to the fulfillment of the reformulated version of Goal 2 (the new desire explanation provided by the participant) is not clear, an additional belief can be provided, but providing the reformulation of Goal 2 (the new desire) is primary to resolving the puzzle.

In response to goal-blocking puzzles like the one just presented, it is of course possible that participants would provide false belief explanations; for example, participants could explain the above example by saying, "The man falsely thought that hanging up on his sister would effectively communicate his love." However, we argue that such puzzles will primarily elicit

desire explanations for two reasons. First, in the context of goal-blocking puzzles, many belief explanations – including false beliefs – will amount to uninformative confabulations about hypothetical means-end relationships (e.g., about how the man believes that his action will, in some unspecified way, fulfill his desire to express love for his sister) with no inferential grounding in the stimulus itself.[6] Because prior knowledge dictates that the goal achieved by the stated action largely contradicts the stated desire, an ad hoc statement that the agent believes otherwise is unlikely to contribute to genuine understanding (See Figure 3.3b).

Secondly, and more fundamentally, we argue that when there is an absence of any clear "destination" at all, as is the case in goal-blocking puzzles, participants' understanding will be enhanced most significantly by providing such a destination. They will clarify this destination – the higher-order goal sought by the agent – by providing one or more desires that subsume the existing goals under a single, hierarchical structure. For example, by responding that "the man wanted to show "tough love" to his sister," a participant provides a new desire explanation (a reformulation of the original "Goal 2" as labeled in Figure 3.3c, the content of which is *to show his love for his sister*) that subsumes the goal of *showing displeasure* (or Goal 1 in Figure 3.3c) under the goal of showing love.

One complication does arise, however. As Figure 3.3c shows, introducing a new or reformulated desire in a hierarchical structure also means introducing a new means-end relationship: this time, showing how the old goal (e.g., expressing displeasure, Goal 1 in Figure 3.3c, which itself is a kind of action description) serves as a means to the achievement of the new desire (labeled "Goal 2, reformulated" in Figure 3.3c). If the new desire is a good 'fit' with the existing goals in prior knowledge, this means-end relationship will be implied (e.g., it is fairly

[6] If, for example, the stimulus offered some compelling background to motivate why the agent might have this particular false belief in this case, the structure of the puzzle would fundamentally change.

obvious that communicating displeasure can be a way to show "tough love"). However, if participants' newly provided desire explanation ("Goal 2, reformulated") is not a perfect fit, an additional belief explanation may be required to show just precisely how the goals are hierarchically related. Crucially, however, even if a connecting belief is required, we predict that it cannot be an effective solution to a goal-blocking puzzle unless it is accompanied by a desire explanation as well. Furthermore, if a desire can resolve the puzzle by itself without additional support from a belief, this is the most parsimonious solution to the puzzle (see Harman, 1965 and Keil, 2006 for a discussion of the role of parsimony in explanations).

In summary, we predict that (I) Goal-blocking puzzles will primarily elicit desire explanations, though they may also elicit belief explanations, and that (II) Means-end puzzles will primarily elicit belief explanations.

3.4 Study 5 Methods

For Study 5, we created puzzling stimulus sentences by pairing actions with puzzling belief or desire explanations for those actions, e.g., "The lawyer placed a call to one of his clients [action] because he wanted to see what the weather was like [desire reason]. Use of these action-reason pairs allowed us to create two different puzzle types. While Goal-blocking puzzles presented two inconsistent goals (one implied in the action, one stated or implied in the reason), Means-end puzzles presented a goal (stated or implied in the reason) that was achieved via an atypical means (the action). While desires explicitly stated a goal (e.g., "because he wanted to…"), beliefs only implied this goal (e.g., "because she knew the supermarket was about to close.") In the final set of items, stimulus reason type (belief or desire) was fully crossed with puzzle type.

Pretesting. A total of 21 stimulus sentences containing actions paired with ill-fitting reasons were pretested on two main dimensions, *goal compatibility* (to select goal-blocking items) and *action commonality* (a test of the commonness with which a particular action is used as a means to achieve a particular goal, to select means-end items). 8 of these items, four in each group and two of each reason type (belief or desire) per group, were ultimately selected for use in the study.

The format of pretest questions varied as a function of the type of explicit reason with which the stimulus was paired (belief or desire). These differences are noted for each dimension. Because of slight differences in the instructions across these item types, items containing belief and desire reasons were pretested between subjects. (See Appendix for complete list of items.)

Goal Compatibility. Consider a belief-containing sentence: "She strapped on her cross-country skis because [she believed][7] the supermarket was about to close." Pretest participants initially read the first part of the sentence, "She strapped on her cross-country skis," and then answered the question "What was her goal?" (e.g., "to enjoy herself in the snow."). Then they read the entire sentence ("Actually what happened was the following…[full sentence]") and were asked, "Now that you've seen the whole sentence, what do you think her goal was?" (e.g., "to get to the supermarket on time"). Then, for goal compatibility, participants rated how *compatible* the initial goal that they had provided was with the final goal on a scale of 0 (Totally incompatible) to 8 (Perfectly compatible). For a desire-containing sentence, e.g., "The woman sent her friend a check for $1,000 because she wanted to pay her back for lunch," the first goal question was the same as for the belief version. The second goal question was omitted, however, because the second half of the sentence (e.g., "she wanted to pay her back for lunch") itself provided the

[7] Even without use of mental state markers such as "he believed," such formulations are still considered by explainers to be beliefs of the agent (Malle et al., 2000). There are, however, of course pragmatic distinctions between the choice to use or omit a mental state marker.

overarching goal of the behavior. This goal content was used as the second goal in the goal compatibility question.

Action commonality. For belief-containing items, participants rated how common it was for a person to perform the initial action (e.g., "strapping on her cross country skis") when trying to achieve the final goal inferred by the participant (e.g., "getting to the supermarket on time") on a scale of 0 (Not at all common) to 8 (Very common). For desire-containing items, participants rated how common it was for a person to perform the initial action (e.g., "A woman sending a $1,000 check") when trying to achieve the final goal specified in the desire contained in the stimulus sentence ("paying the friend back for lunch").

General puzzlingness. In addition to the individual puzzle types, we also pretested for the general puzzlingness of each stimulus sentence in a third set of subjects. This question was worded as an "understanding" question: "Given this explanation, how well do you *understand* this action?" and "puzzlingness" ratings were created using the formula $P = 8 - U$.

Stimulus selection. Four items were selected for inclusion in the means-end category, and four in the goal blocking category. For each dimension, items were eligible for selection if they were below their respective scale midpoints (4). Goal-blocking items had a mean of $M = 2.02$ ($SD = 0.45$) on the goal-compatibility question, and Means-end items had a mean of $M = 2.72$ ($SD = 0.98$) on the action commonality question.[8] In addition, while item selection was based on a single-question criterion, participants responded to both questions for each item. Thus a single composite score for each item could also be calculated by subtracting that item's action commonality score from its goal-compatibility score; on this composite score, a more positive

[8] Despite several rounds of pretesting, it was not possible to distinguish the categories along both dimensions. In a test that included 6 out of the 8 selected items, the two categories differed on the goal compatibility dimension, $t(69) = 3.4$, $p < .01$, but not on the action commonality dimension, $t(69) = .58$, *ns.* The consequences of this asymmetric differentiation are detailed in the analysis section.

score indicated higher goal compatibility than goal commonality, while a negative score indicated the reverse.

Task and design. The task was identical to that presented in Studies 1, 2, and 4 of Chapter I. Participants read the stimulus sentences one by one and were invited to "add whatever information you think is needed for the sentence above to make better sense." The eight items were presented to participants in two forms, consisting of one pseudorandom order and its reverse.

Participants. Seventy-nine participants (39 female) completed the task on Amazon Mechanical Turk in exchange for monetary compensation. Their average age was 34 years, and 51% had a four-year college degree or higher level of education.

Data Treatment. Two coders (one blind to condition) were involved in the classification of 632 explanations using the F.Ex coding scheme (Malle, 1998/2004), which classifies mental state explanations into belief and desire reasons, as well as distinguishing between mental state and non-mental state (causal history of reason) explanations (see Malle, 1999, for more details and evidence on this distinction). For reliability purposes, the two coders each classified responses to 128 items. Coders reached 100% agreement on whether a statement was codable as an explanation (K = .68), and 96% on mental state reasons versus causal history explanations (K = .69), and 99% agreement on belief versus desire reasons (K = .86). Both coders independently classified a second group of 437 responses; disagreements were resolved in discussion. A final 64 responses were then classified by a single coder.

3.5 Results and Discussion

Aggregate analysis. We first aggregated the number of explanations provided by each participant over category, such that each participant received a mean score for beliefs and desires

for Means-end puzzles and Goal-blocking puzzles (a total of four scores). A 2 (Goal-blocking vs. Means-end puzzle type) X 2 (Belief vs. Desire reason) repeated measures ANOVA revealed a significant interaction, $F(1, 70) = 31.02$, $p < .001$. That is, participants provided more beliefs in response to Means-end puzzles ($M = 0.86$, $SD = 0.47$) than in response to Goal-blocking puzzles ($M = 0.59$, $SD = 0.47$), while they provided more desires in response to Goal-blocking puzzles ($M = .71$, $SD = .44$) than in response to Means-end puzzles ($M = 0.39$, $SD = .32$). There was also a main effect of explanation type, with participants providing more belief ($M = .74$, $SD = .34$) than desire explanations overall ($M = .54$, $SD = .31$), $F(1, 70) = 5.29$, $p = .02$.

Disaggregated analysis: Multilevel Poisson regression. In response to each item, participants provided as many explanations as they deemed necessary to make sense of the sentence. In response to every item, then, participants could provide one or more beliefs and desires (they were not limited to one or the other type of explanation). Because individual, disaggregated scores represented count data (number of beliefs and desires given per item), we analyzed the data using a Poisson distribution, a commonly used distribution for (disaggregated) count data, as ordinary least squares regression is generally a poor choice for such data (O'Hara & Kotze, 2010).[9] To examine the effects of individual items' pretested goal-blocking and means-end ratings on specific rates of beliefs and desire explanations given, we initially performed a multilevel Poisson regression for the number of beliefs and desires given by each participant in

[9] The data were somewhat underdispersed ($M = 0.66$, $SD = 0.43$, Deviance/$df = .78$) compared to the Poisson distribution, which assumes the mean and variance are equal. We thus examined an alternative model as well. Since the modal values in the count data were "0" (no explanations of a given type) and "1" (one explanation of a given type), transforming the response variable (number of explanations of each type) into a binary (either providing an explanation of that type, "1," or not providing any explanations of that type, "0") provided an alternative approach. This did lead to some loss of information, but it allowed us to perform a multilevel logistic regression, which requires less stringent distributional assumptions. This model was distinguished from the Poisson model primarily by a significant improvement in model fit when a random slope for number of belief vs. desire explanations by participant ID was also included ($\chi^2(3) = 32.52$, $p < .001$). While this model led to somewhat different values for regression coefficients**,** the overall interpretation of fixed effects in the model did not differ from the Poisson model, nor did it differ from a version of the logistic model run without random effects parameters.

response to each of the 8 items (for a total of 16 data points per participant). An analysis of random effects revealed that 0% of the variance in number of explanations provided was explained by including a unique intercept for each subject. In addition, inclusion of random slope for number of explanations in each category (belief or desire) by subject also did not improve performance of the model, $\chi^2(3) = 1.94$, *ns*. With no significant contributions from random effects, the analysis proceeded with the consideration of fixed effects only.

Each item's pretested rating for Action commonality and Goal compatibility was entered into the model as a continuous predictor, as was that item's combined Goal compatibility - Action commonality score. In addition, the general degree to which each item was puzzling was included as a continuous predictor.

The combined pretest ratings (Goal compatibility – Action commonality) alone did not significantly predict the number of belief or desire explanations participants provided**.** In the final model, goal compatibility emerged as a significant predictor for both belief and desire explanations; for every increase of one unit in goal compatibility, participants gave 22% more belief explanations ($p < .01$, 95% CI, 7% to 39%) and 36% fewer desire explanations ($p < .001$, 95% CI, a 21% to 48% decrease). In addition, while action commonality did not emerge as a significant predictor for either explanation type, general ratings of the "puzzlingness" of an item did. Participants increased the number of belief explanations they provided by 23% ($p = .01$, 95% CI, 5% to 37%) and decreased the number of desire explanations by 25% ($p = .04$, 95% CI 2% to 54%) for every unit of increase in the item's puzzlingness.

Why might puzzlingness ratings have superseded action commonality in the prediction of participants' providing beliefs? One notable point is that action commonality ratings were actually quite similar across Means-end and Goal-blocking items ($M = 2.72$, $M = 2.25$,

respectively) so the two groups of items were much better distinguished by the goal-compatibility dimension than the action commonality dimension. Because in the present pretests, action commonality and puzzlingness ratings were collected in separate groups of subjects, we could not compare their covariances between these two dimensions directly. However, a recent set of pretests drawing on the same theoretical framework (DiGiovanni, Korman, & Malle, 2016) elicited both action commonality and puzzlingness ratings for each item. Similar to the current data, the results of an ANOVA for these earlier data also showed no initial difference between Means-end and Goal-blocking items on action commonality ratings, but when puzzlingness ratings were entered as a covariate, the two groups of items turned out to differ significantly on action commonality. This occurred because Means-end items were overall less puzzling than Goal-blocking items and, once this difference was accounted for, the items differed on action commonality. While not definitive for the present study, this finding suggests that the failure of action commonality to directly predict participants' belief usage may have been due to the fact that means-end items are less puzzling than goal-compatibility items overall.

3.6 General Discussion and Future Directions

What are the primary determinants of the choice between beliefs and desires? Our strongest finding was that, the more incompatible the initial goal implied by the action was with another goal the agent had in performing that same action, the more likely participants were to provide desires, and the less likely they were to provide beliefs. This confirms the prediction that, when an action lacks a clear goal, people will search, first and foremost, for the "final destination" of that action, providing desire explanations. However, in response to goal-blocking puzzles, participants did not cease to give belief explanations entirely. In fact, although they provided fewer beliefs in for Goal-blocking puzzles than for Means-end puzzles, participants still

provided more beliefs for goal-blocking puzzles than they did desires for means-end puzzles, $t(70) = 2.76$, $p < .01$). This suggests that, in keeping with the predictions outlined in Figure 3, beliefs may also be an important, if secondary, component when responding to goal-blocking puzzles. In contrast, in keeping with the prediction that desires are a largely uninformative and inefficient way to resolve means-end puzzles, participants gave very few desires in this case (less than half as many desires as beliefs).

Two limitations of the present study should be noted. First, the use of only eight unique items in total limited the variability of our items along the two main pretest dimensions. Future studies with a larger number of items and a more variable range of scores over the two dimensions should replicate the present finding. Secondly, while the use of actions paired with puzzling reasons as stimuli did supply sufficient variation to differentially elicit beliefs and desires as a function of puzzle type, these stimuli in particular, both in the present study and in a previous study (Chapter 2, Study 3, this dissertation, $F(1, 39) = 7.20$, $p < .01$), elicit a greater number of belief than desire explanations overall. This prevalence of beliefs stands in contrast to much of the literature on action understanding, which suggests that desires are the much more frequent inference. Thus, if the current findings are to be generalized to settings of more naturalistic human action understanding, they should be replicated with more naturalistic behaviors.

In spite of these limitations, our study does make a distinct contribution to the study of the roles of belief and desire in action understanding. Previous work examining the role of both mental states within a single study is limited. Such work generally explores the distinct roles of these two mental states by presenting an action that is primarily physical in nature, and queries inferences using forced-choice or probabilistic measures (e.g., Baker et al., 2014; Richardson et

al., 2012). For example, in Richardson et al. (2012), child and adult participants made belief and desire inferences in response to a scenario in which a bunny had a preference for one of three fruits that had fallen from a tree. In several conditions, the bunny either approached the fruit that was in plain sight, or passed by one fruit in search of other fruits he presumed (but did not know for sure) were hidden behind a wall. While this study can examine participants' abilities to infer the relative preferences of the bunny on the basis of his actions (for example, if he passed by a fruit in plain sight to look for other fruits behind the wall, he clearly did not prefer the first visible fruit), it represents a highly constrained context in which the basic preference of the agent in question (to obtain fruits) is already provided to participants; it is only the specific content (which of three fruits) that the participant needs to infer. And while this work does require participants to represent the agent's belief, it is a simple perceptual belief about objects that is very easily determined from the configuration of the physical environment. The current study was distinct from this earlier work in two important respects. First, similar to explanations for actions in the real world, the contents of mental representations were limited only by participants' prior knowledge – not by the tight constraints of an agent moving in a physical environment. In order to resolve the puzzles, participants had to construct rich representations of the agent's beliefs that moved beyond simple facts of the agent's perceptual access. And crucially, in contrast to the previous work, we did not provide the agent's basic goal in our goal-blocking puzzles. Participants had to do more than "fill in" the content of the agent's desire (e.g., by choosing fruit A, B, or C): solving the puzzle required generating a novel inference of the agent's goal. Second, in contrast to the previous use of forced-choice measures, we elicited participants' open-ended, self-generated representations of agents' mental states. This methodology reveals that the apparent differences in structure between belief and desire in action

explanation -- suggested by only a small number of earlier studies in adults -- is not the mere consequence of a constrained response medium.

CHAPTER 4.

Action Understanding in High-Functioning Autism: The Faux Pas Task Revisited

4.1 Introduction

4.1.1 Autism and Theory of Mind: Belief and Action Understanding

Autism spectrum disorder (ASD) is a developmental disorder characterized by deficits in reciprocal social interaction and communication (DSM-V, American Psychiatric Association, 2013). In contrast to typically developing children, autistic children are widely described as having deficits in "theory of mind," or the ability to represent the mental states of other people. More than thirty years of research has documented young' children's emerging ability to represent another person's false belief (Wimmer & Perner, 1983; see Wellman, Cross, & Watson, 2001, for a review) around the normative age of approximately four years. Although theory of mind deficits in autism spectrum disorders were originally conceived solely in terms of ASD children's early failures to pass false belief tasks at this normative age (Baron-Cohen, 1985; Frith, Morton, & Leslie, 1991; Leslie & Thaiss, 1992), more recently questions about the abilities of autistic children and adults to competently represent a range of other mental state concepts – such as desires, intentions, and emotions – have come into view (Hamilton, 2009). Studies of these other concepts in individuals with autism spectrum disorder are less definitive. For example, a small literature has grown around the idea that, like typically developing 18-month-olds (Meltzoff, 1995), autistic children may also understand that observable actions are caused by an agent's unobservable intentions. When instructed to do so, preschoolers with autism are able to imitate object-directed actions with a clear end-state and will act out the physical outcome behind an incomplete action, even if they have not seen it completed

beforehand (Carpenter, Pennington, & Rogers, 2001; Aldridge, Sweeney, Stone & Bower, 2000, Berger & Ingersoll, 2014). However, they have more difficulty imitating non-meaningful gestural actions that do not result in concrete physical changes in the world (Vivanti Nadig, Ozonoff, & Rogers, 2008), and unlike typically developing (TD) children, they do not imitate information about the means used to achieve a particular goal (e.g., achieving the goal of flipping a switch using either a *motion of the hands* or a *motion of the feet*) with precision (Hobson & Lee, 1999). In contrast to their ability to understand an imitate physical outcomes of intended actions, children with autism struggle to do so when outcomes are abstract or the cues to intention are social rather than physical (Berger & Ingersoll, 2014). In summary, even though children on the autism spectrum have a basic understanding that purposeful bodily movements are caused by the agent's underlying intentions, they may be less sensitive to the structural features of more complex actions, such as the precise relationship between means and goal (Tomasello, et al., 2005).

In addition, in contrast to their early struggles with false belief understanding, children with autism grasp the concept of desire at the same mental age as typically developing children, demonstrating an understanding of desires as inner states that cause behavior (Peterson et al., 2005; Tan & Harris, 1991). However, a recent study also suggests that young children with ASD may lack a more sophisticated understanding of the subjectivity of desire, failing to correctly predict another person's behavior when that person's desire conflict's with the child's own (Broekhof et al., 2015).

Mental state concepts such as intention, desire, and belief do not exist in isolation, however. They make up a framework of interrelated theory of mind concepts that is intimately yoked to human action understanding and the full-blown representational concept of intentional

action: that an action is only performed *intentionally* if the agent had a desire for the action's outcome and a belief that her action would lead to – serve as a means to achieving – that particular outcome (Malle & Knobe, 1997; Kashima et al., 1998; Reeder, 2009). Beyond the mental state concept of intention (Searle, 1983), which is solely a decision to act, the concept of intentional *action* links agents' mental states with their causal background as well as with the causal consequences of behavior. In the mature, typically developing adult, each of these concepts is intimately yoked to humans' ability to interpret and explain human behavior. In the present chapter, I explore whether high-functioning adults on the autism spectrum have an intact concept of intentional action: one that allows them to understand even complex social behaviors that require the social perceiver to consider multiple components of intentional action.

4.1.2 High-functioning ASD Adults: From Belief to Intentional Action

By the time they reach adulthood, many high-functioning adults on the autism spectrum are capable of passing traditional false belief tasks as well as "second-order" versions of such tasks (in which one character's false belief represents another character's belief) (Bowler, 1992). Because ASD individuals require higher verbal ability than typically developing adults to pass such tasks, researchers have proposed that ASD individuals rely more than TD individuals on the representational power of language to pass such tasks via deliberate, conscious calculation (Happé, 1995), and the reliance on specific features of language, such as complement syntax (Lind & Bowler, 2009). However, the persistence of social deficits in high-functioning autistic adults in spite of eventual ToM task passage (e.g., Klin, 2000, Klin et al., 2003) suggests that such tasks do not capture these social deficits' core features. Hence, autism researchers developed more "advanced" theory of mind tasks to highlight the persistence of theory of mind

deficits – broadly conceived – in more naturalistic settings. Although researchers have succeeded at demonstrating that individuals on the autism spectrum struggle with these novel tasks (Baron-Cohen, 1999; Zalla et al., 2009; Happé , 1994, Begeer, Malle, Nieuwland, & Keysar, 2010), the precise mechanisms underlying these struggles have gone largely unexplored. In the present chapter, I first introduce two of the major "advanced" theory of mind tasks used by researchers. Then I report on two studies that use a modified version of one of these tasks to explore competence in ASD adults' understanding of the concept of intentional action.

Beyond requiring a simple ability to represent belief, these more advanced theory of mind tasks present ASD individuals with complex intentional actions. While understanding that another person can have a desire distinct from one's own desires is sufficient for a basic representational concept of desire, and understanding that a person can hold something to be true even if it is in fact false is sufficient for a basic concept of false belief, intentional action understanding involves additional components: One must not only grasp each of the component mental states but also the complex interrelations between the specific beliefs and desires on the agent's mind, the actions he performs, and the causal consequences of those actions. Only then is the meaning of the outcome revealed: the social perceiver either grasps the specific intention underlying the action that brought about that outcome, or the circumstances that explain how that outcome was brought about unintentionally.

One such 'advanced' theory of mind task is the Strange Stories task (Happé , 1994). This task presents high-functioning ASD individuals with non-literal utterances, a task that was designed to be more "contextually embedded and realistic" than traditional first- and second-order false belief tasks. One "strange story" depicts a lie, in which a girl named Anna mistakenly knocks over and breaks her mother's crystal vase. When Anna's mother returns

home to find the broken vase, Anna says, "The dog knocked it over, it wasn't my fault!" Participants are asked (1) whether what Anna told her mother was true, and (2) why Anna said this. A fully correct explanation of the action needs not only to recognize that Anna's statement is false, but also that she makes the false statement with a particular communicative intention: she wants her mother to believe that she was not responsible for breaking the vase, and she thinks that deflecting blame to the dog will convince her mother (a belief that her statement is a means to the fulfillment of her desire). The task thus presents a potential challenge to participants not only because it requires a recognition of the falsity of Anna's statement in the given context, but because of the effect that Anna intends her statement to have on her mother's mental state. While autistic adolescents and young adults used as many mental state terms as control participants, they often failed to identify the mental state *appropriate* to the situation at hand – those that cite not just any thought that Anna or her mother might have had, but ones that correctly identify the particular belief or desire behind Anna's communicative intention. False belief tasks successfully demonstrate conceptual competence by revealing a clear distinction between the representation of epistemic states and the representation of reality (Dennett, 1978; Premack & Woodruff, 1978). But rather than simply decoupling belief from known reality, the Strange Stories task decouples the surface interpretation of an utterance from the specific intention underlying that utterance: the planned effect that the utterance will have on another person.

However, because the Strange Stories task does not describe how Anna's mother responds to her statement, its focus is primarily on the mental state of intention – the thoughts and plans Anna had that drove her decision to make such a statement – rather than on the eventual impact of her statement once it was made. As such, the task does not fully measure

understanding of the concept of intentional *action*, which starts with the mental state of intention and its constituent concepts but ends with the realization or non-realization of the intention. For example, Anna may have intentionally convinced her mother of her innocence or unintentionally made her mother angry because of the lie.

In contrast, the "faux pas" task (Baron-Cohen et al., 1999; Zalla et al., 2009) explores participants' understanding of *both* the mental states leading up to an intentional action and the outcomes it causes. This task describes a context in which one character (the speaker) makes a statement that is unintentionally offensive to the listener because the speaker has a false belief. For example, in one story, Jane moves into a new apartment and purchases new curtains for the windows. When her best friend Lisa comes over, she says to Jane, "Oh, I hope you're going to get new curtains! These ones are awful!" Lisa's comment is offensive to Jane, but when asked why Lisa said that, typically developing individuals infer that Lisa did not know the curtains were picked out by Jane herself. In contrast, while individuals on the autism spectrum can detect that something was "wrong" with Lisa's comment, they struggle to detect that Lisa made the comment unintentionally – because she had a false belief, and believed the statement would fulfill a positive or neutral desire (e.g., to be helpful with decorating). The outcome is not just an unfortunate side effect of an otherwise fulfilled intention; the complete falsity of the agent's belief actually precludes the fulfillment of the speaker's desire. High-functioning adults on the autism spectrum are able to detect that someone said "something awkward," but they demonstrate a mixed pattern of responding in their detection of the character's false belief and positive desire. Sometimes, they acknowledge that the speaker had a positive desire, but fail to correctly infer the speaker's belief state, while in a small number of cases (10% of all responses),

they even incorrectly attribute a negative intention to the speaker (e.g., Lisa wants to insult Jill's taste in décor) (Zalla et al., 2009).

4.1.3 Intentional Action Understanding: Integrating Mental States, Action, and Outcome

Together, the strange stories and the faux pas task suggest that, even while ASD adults have come to a competent basic understanding of beliefs in highly structured tasks through the deployment of compensation strategies, they may still lack the ability to appropriately link beliefs, desires, and actions together to achieve a full conceptual understanding of behavior and its consequences in the world. But besides a broad failure to fully grasp the intentional action concept, there is another possible source of deficit: a more specific failure to properly infer the agents' mental states. Only in the presence of explicit mental state information – which is then integrated by participants – can this hypothesis be ruled out.

An emerging literature has partly addressed this issue by explicitly presenting mental state information to participants in the context of eliciting moral judgments of an agent's behavior. In the presence of such information, there is no need for participants themselves to draw inferences to generate mental states. For example, Channon et al. (2011) presented participants with intentional and unintentional behaviors for which the main character's intention (or lack thereof) was clearly stated (e.g., "he purposefully gave his wife an overdose" or "his short-sighted wife gave him an overdose by mistake"). They then elicited blame judgments, and found that the blame judgments of ASD adults were *more* sensitive (in the expected direction) to the explicitly provided intention information than were those of typically developing control participants. This finding suggests that ASD adults have an intact ability to integrate mental state information for moral judgment. However, in addition to aiding participants by providing

mental state information explicitly, Channon et al.'s task was actually simpler in structure than the earlier-described advanced theory of mind tasks. Rather than describing a longer sequence about an agent who possesses certain mental states, performs an action on the basis of those mental states, and brings about an outcome, Channon's task provides an action description in which the mental states are *present* in the case of an intentionally caused outcome ("purposefully"), and simply *absent* in the case of an unintentionally caused outcome ("accidentally"). Participants could thus solve the task with the simple knowledge that "accidental" negative actions are less blameworthy than "purposeful" ones, without having to integrate any specific desire or belief with any specific piece of outcome information to determine *which action* in the story was performed intentionally or accidentally. In other words, Channon et al.'s task does not address the more complex case in which an agent may have performed *some* action intentionally, but that action may have had *also* had unintentional outcomes. That is, the agent did have *some* relevant mental states leading up to the behavior, but these mental states may be inconsistent with the outcome that ultimately results.

In contrast, in Moran et al. (2011), mental states, actions, and outcomes were all described separately, and participants had to actively integrate each of these components in order to make an appropriate blame judgment. Moran et al. (2011) found that, given this more complex structure, even individuals on the autism spectrum with intact belief understanding judge actions that lead to negative outcomes but are brought about by misguided beliefs to be less morally permissible than do typically developing adults. In spite of receiving explicit belief information, high-functioning adults with ASD still *weighted* these explicitly stated beliefs less heavily (and the outcomes more heavily) in their moral judgments than did typically developing adults.

While this result suggests a possible deficit in the integration of inconsistent mental state information with the explicit outcome of an intentional action, it is not definitive for several reasons. First, because Moran et al. elicited *permissibility* judgments rather than blame judgments, it is also possible that ASD participants' permissibility judgments, which are essentially judgments of the normative acceptability of an action, would diverge from their blame judgments, which more straightforwardly entail the consideration of the agent's mental states (Malle, Guglielmo, & Monroe, 2014). In addition, it is notable that while the scenarios were described as providing information about the neutrality or negativity of the agent's intentions, the authors described only the agent's belief state, leaving it to participants to use that belief information to infer that the agent had a neutral or negative desire in the situation. And finally, even though ASD adults rated the neutral-intention/negative-outcome action as less morally permissible than did typically developing participants, they still rated this action as being *more* morally permissible than the negative-intention/negative-outcome condition, a finding that qualitatively suggests at least some preserved ability to integrate mental state information about actions that result in inconsistent outcomes.

Even if we accept documented differences between ASD and typically developing participants as indicative of such a deficit, integrating mental states with outcome information for intentional action understanding may be quite distinct from doing so for moral judgment. In both cases, integration requires the reconciliation of conflicting mental states and outcomes – often, ones that point to oppositely-valenced conclusions – to reach a single interpretation of the blameworthiness or meaning of the action. In the case of moral judgment, however, ASD individuals struggle with mental state integration in only one case: the case of accidental harm. In contrast, in the case of attempted harm, ASD individuals perform comparably to controls.

Like the case of accidental harm, the case of attempted harm involves mental states that conflict with outcomes: the agent has a negative intention but the action leads to a neutral outcome. What differs is that in the attempted harm case, there is no harmful outcome. And some authors (e.g., Cushman et al., 2013) have argued that such cases are not fundamentally tied to the typical development of moral concepts: intent-based moral judgment develops when children increasingly integrate intention information with an early-developing process of harm-based moral judgment. Similarly, Malle, Guglielmo, & Monroe (2014) argue that the process of moral judgment is triggered by the detection of a norm-violating event. Therefore, if ASD individuals appear to weigh outcomes more heavily than mental states than do TD individuals for moral judgments of accidental harms, this provides evidence only for a struggle with the integration of mental state information for the specific, harm-based process of moral judgment, and not necessarily for the more general integration of mental states with outcomes. In other words, it is possible that a process less yoked to the detection of harmful outcomes would yield a different profile of integration abilities.

Intentional action understanding may well be such a process. When the mental states behind an action conflict with the outcome of that action (e.g., a person intends to make a mean remark, and instead inadvertently pays a complement, or vice versa), the meaning of the action – the description of which action was intentional, and the outcomes that were brought about unintentionally – conceptually depends *both* on the mental states of the agent, and on the outcome of the action, *regardless* of whether the outcome is harmful or norm-violating. Thus, while moral judgment involves *first* detecting a harm and *then* taking into account (integrating) the agent's mental states, the integration of mental state information for intentional action understanding requires the social perceiver to consider the consistency of each component –

mental states, observed action, and outcome – with each other component, and then to resolve these inconsistencies by reaching one, integrated representation of which action was intentional, which outcomes were brought about intentionally, and which outcomes the agent caused, but did not intend to bring about.

However, even if intentional action understanding is distinct from moral judgment in typically developing children and adults, it is still possible that these two processes are conflated in autism, with normative judgments playing a strong role in the conceptual understanding of intentional action. If this is the case, then the judgment of an action's outcome as in violation of a normative rule – such as judging Lisa's comment about the curtains as being in violation of the rule, "one should never insult a friend's decorating choices," – could preclude the consideration of conflicting mental state information, and drive ASD adults' overall conceptual representation of intentional action. Zalla and colleagues (Zalla & Leboyer, 2011; Zalla et al., 2009) have argued that this is a primary finding of previous work on the faux pas task. Specifically, they argue that while moral considerations appear play some role in intentional action understanding in typical development (Knobe, 2003), they may play an even greater role in ASD individuals' later-developing ability to understand such actions. Specifically, according to Knobe (2003), typically developing adults judge a foreseen morally bad side effect as more likely to be brought about intentionally than one that is morally good. Zalla et al. (2009) argue that ASD individuals' understanding of such actions may, because it develops later and via some (underspecified) alternative developmental path, be driven by such moral considerations even when the outcome of the action is *not* foreseen. Such a conceptual model of intentional action on the part of ASD individuals – one that attributes negative intentions on the basis of negative outcomes with little regard for mental states – can partly account for previous findings that members of this

population occasionally appeal to negative intentions in their behavior explanations of actions that result in faux pas. However, even if these explanations are more common among ASD participants than in controls, they are still relatively rare in ASD individuals. Furthermore, this account cannot explain the large number of correct mental state explanations that ASD individuals are still able to provide.

Thus, even if individuals with ASD are capable of inferring a character's individual mental states, it is still unknown whether individuals on the autism spectrum have an intact concept of intentional action that enables them to integrate mental state information with information about actions and outcomes that are causally linked to these mental states. That is, even if they are able to infer what the person's belief state and desired outcome were, they may be unable to understand how the person's mental states and actions causally relate to each other and to an outcome that is neither desired nor foreseen by the agent in performing the action.

4.2 The Faux Pas Task: Two Accounts of Inferential Deficit

In the following sections, I will present a modified version of the faux pas task that explores ASD adults' understanding of intentional action by explicitly presenting mental state information to participants. However, before proceeding to the modified version of the task, it is important to address what this task measures in its standard form. In the literature, the faux pas task is treated as an advanced 'theory of mind' task, often used as a benchmark for measuring theory of mind capabilities of adolescents and adults with autism (e.g., Banerjee & Watling, 2005; Bottiroli et al., 2016). However, it is not clear that this task is simply a more advanced way of revealing the enduring theory of mind deficits that ASD adults have had from childhood (henceforth referred to as the ***enduring mental state inference deficit*** account), or whether it

simply presents a case in which, for identifiable reasons, cognitive compensation strategies that succeed on structured first- and second- order false belief tasks are ineffectual (referred to as the ***knowledge-based inference deficit*** account).

How do these compensation strategies work, and why might they be ineffectual on the faux pas task? Children with autism start out, of course, by failing false belief tasks at the normative age of four years, and do not come to pass them until at least a verbal mental age of 9 (Happé, 1995). However, a few studies have suggested that children with autism fail false belief tasks but pass structurally similar tests of counterfactual reasoning because false belief tasks require "generativity," or supplying a novel premise not provided in the question (Peterson & Bowler, 2000; Riggs et al., 1998), while counterfactual reasoning tasks do not. This early facility on a structural analog of the false belief task may also shed light on how high-functioning individuals eventually, by early adolescence, come to pass false belief tasks themselves. They may be able to do so precisely because they have by then acquired the kind of general background knowledge that supplies the missing premise in these simple cases of perception, e.g., "perception leads to knowledge." However, this compensation strategy may fall short on faux pas stories, which reach far beyond the relationship between perception and knowledge, and depict a range of social relationships, interactions, and contexts. Such stories may often require the participant to infer a richer array of background information *before* determining the character's mental state. For example, consider one of the original faux pas stories used by Baron-Cohen et al. (1999):

> Kim helped her [mother] make an apple pie for her uncle when he came to visit. She carried it out of the kitchen. "I made it just for you," said Kim. "Mmm," replied Uncle Tom, "That looks lovely. I love pies, except for apple, of course!"

In order to respond to this story correctly, the reader has to infer that Kim's uncle was *not* present when Kim made the pie, even though the story ("when he came to visit") is ambiguous about this fact. In order to correctly infer that Uncle Tom does not know that the pie is apple, the reader must reach this inference by appeal to the general knowledge that (for example) instances of uncles deliberately saying mean things to their nieces are relatively uncommon, and that this is an improbable interpretation of the present scenario. Only when this general knowledge is applied is the inference about Uncle Tom's absence during the pie-baking possible, and only in light of *that* inference is the subsequent mental state inference possible: his lack of knowledge about the pie's actual contents (or false belief that the pie was filled with any one of various other fruits).

It is thus not clear whether the apparent failures of mental state *inference* on these tasks are due to a more general and enduring deficit in mental state inference in more complex social situations, or to a difficulty in the ability to appropriately apply background knowledge to the generation of novel mental states. Indeed, a small literature does suggest that beyond "theory of mind" deficits, individuals on the autism spectrum may struggle to draw productively on their existing knowledge of common social event schemas (Trillingsgaard, 1999; Loth et al., 2008; Loth et al., 2011), failing to identify relevant information in their environment and to recognize probable events in such common schemas. This is an impairment that may well influence their ability to draw mental state inferences: For example, to an individual with ASD, a young child's

uncle may simply be a parent's sibling who sometimes attends family gatherings; the relevance of the fact that he is her uncle – and specifically, the idea that uncles generally tend not to offend their nieces' attempts at hospitality – may never occur to the person with ASD. Rather than reflecting any specific conceptual deficit in intentional action understanding per se, the struggles of ASD individuals to understand complex intentional actions such as those depicted in the faux pas task may thus actually be revealing one of two distinct types of *inferential* deficits: either an enduring inability to extract mental states from any complex action (strictly a "theory of mind" deficit that compensation strategies are unsuccessful at mitigating), or a specific struggle to draw rich, knowledge-based mental state inferences from sparse behavioral information. I address these competing interpretations in the description of the experimental task below.

4.2.1 Enriching the Faux Pas Task: Inference and Integration

Inference. We constructed novel vignettes based on the faux pas task (Zalla et al., 2009; Baron-Cohen at al., 1999) to explore the ability of high-functioning ASD adults to use the concept of intentional action to integrate a character's mental state information with information about the character's behavior, its background causes, and outcomes. In the "No Information" condition, participants were presented with no explicit mental state information, as in the original Faux Pas task. In three additional conditions, they were presented with either the character's belief only, the character's desire only, or the character's both belief and desire (the "Full Information" condition). A sample story we used in the Full Information condition follows. In the "No Information" condition, the text in bold is omitted:

Clara is very short and dresses plainly. One day she goes to pick up her son James from school early for a medical appointment. Clara enters the school and spots James's teacher, Mrs. Hayes. **Mrs. Hayes thinks that Clara is a student lost in the hallway. [Belief] Mrs. Hayes wants to help [Desire].** Before Clara can ask after James' whereabouts, Mrs. Hayes looks at Clara and says, "Have you lost your class, honey?"

Every condition, including the No Information condition, contained background information that would afford a belief inference with minimal prior knowledge added to explicit story information such as "Clara is very short and dresses plainly." In addition, if the original faux pas stories provided any affordances for inference at all, these were often affordances for the character's *lack* of knowledge rather than the background for the actual *content* of the character's mental state. In contrast, our stories provided background that could serve as a direct affordance for inferring this mental state content. If participants continue to struggle to produce belief inferences in the No Information condition in spite of these additions of inference-ready background information, this may indicate that the original faux pas stories do indeed measure an 'advanced' theory of mind capacity, or at least some capacity that is independent of the background knowledge needed to draw appropriate inferences.

Integration. Regardless of whether the original faux pas task measures "advanced theory of mind" or the application of general knowledge, faux pas stories present an apt method to explore more subtle aspects of intentional action understanding because they present an utterance motivated by a benign intention (e.g., a desire to help) and a false belief (e.g., about the identity of the person in the hallway) that results in an unintended outcome (an utterance that takes on

meaning for the hearer that was unintended by the speaker, resulting in an offense of the listener and an "awkward" situation for the speaker). This outcome is also not a merely undesirable side effect: it actually *precludes* the fulfillment of the desired outcome (e.g., instead of being helpful, or, in some stories, paying a compliment, the speaker has delivered an insult which is neither helpful nor complimentary). As in more sophisticated tests of intention understanding given to young children (e.g., Baird & Moses, 2001), these tasks present actions in which the behavior, its motivating mental states, background causes, and outcomes must each be considered separately. A complete understanding of intentional action is demonstrated when each of these components is included in one integrated representation: a story about *which action was intentional*, and then auxiliary causal explanations for why (and how) unintended outcomes may have also resulted. Appropriately constructed, the faux pas task invites participants to evaluate various aspects of the action's meaning, outcome, and emotional impact in both forced-choice and open-ended formats.

By presenting the agent's mental states explicitly in some items, we can examine two primary integration tasks. First, the social perceiver must integrate mental state information with the surface description of the action to devise a novel, accurate description of *which action* was intentional (*mind-action* integration). For example, if individuals on the autism spectrum are simply inclined to regurgitate information given to them, they may parrot back mental state information when it is provided but fail to use this information to produce a novel action description (e.g., correctly noting that "Lisa is trying to help with the curtains"). This task requires the linkage of both belief and desire information to the same intentional action. Mind-action integration is depicted in Figure 4.1.

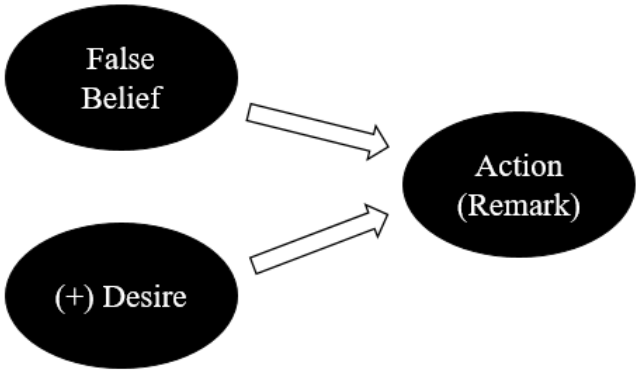

Figure 4.1. *Mind-action integration* takes place when the social perceiver causally links both the agent's false belief and her positive or neutral desire to the action.

The second, related task, *mind-outcome* integration**,** involves identifying outcomes that were *not* caused intentionally (typically not even desired) and causally linking the mental states the agent *did* have to these unintentionally caused outcomes ("Jane is offended by Lisa's remark, which Lisa only says because she falsely believes the curtains to be old."). If integration of mental states and outcomes fails, ASD individuals may see it as perfectly plausible that the speaker has some false belief as a general part of her belief corpus, but may, when explaining the speaker's utterance, still incorrectly generate a true belief (e.g., "She knew her remark would be insulting"). Mind-outcome integration is depicted in Figure 4.2.

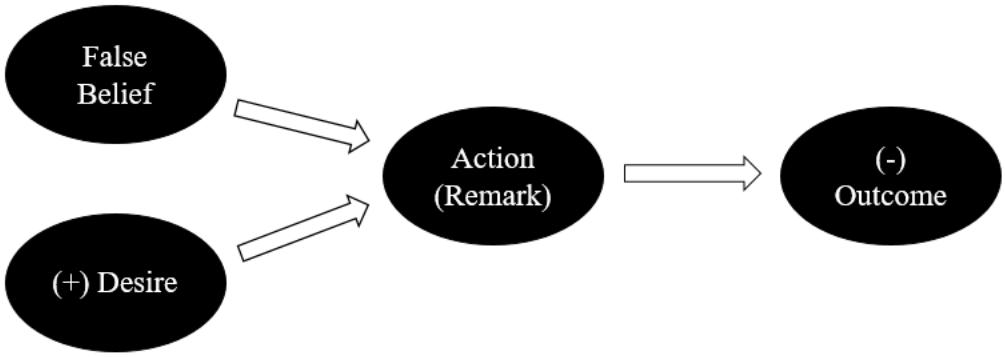

Figure 4.2. *Mind-outcome integration* links the (negative) outcome of the action to the agent's false belief and positive or neutral desire. Although it focuses on the explanation for the outcome and the mental states' role in causing that outcome, the causal path by which this occurs also involves the action.

Finally, beyond the linkage of mental states, actions, and outcomes, the social perceiver can also explain the unintentional outcome by appeal to background causes for the agent's mental states and action; most centrally, to facts of which the agent was not aware (e.g., "She didn't know that Jane had bought the curtains herself.") This lack of knowledge can play a part in producing the agent's mental states (e.g., "That's why she thought her remark would be a helpful decorating suggestion,"), but it can also bypass the agent's intention entirely, providing a more direct causal explanation for why the unintended outcome occurred (e.g., "She didn't know that Jane bought the curtains herself, and that's why her action had this unintended consequence."). This explanation also mainly denies the intention to have brought about the negative outcome. Because background information can be linked directly to both action and outcome, I refer to both of these linkages together simply as *background integration*. Background integration is depicted in Figure 4.3.

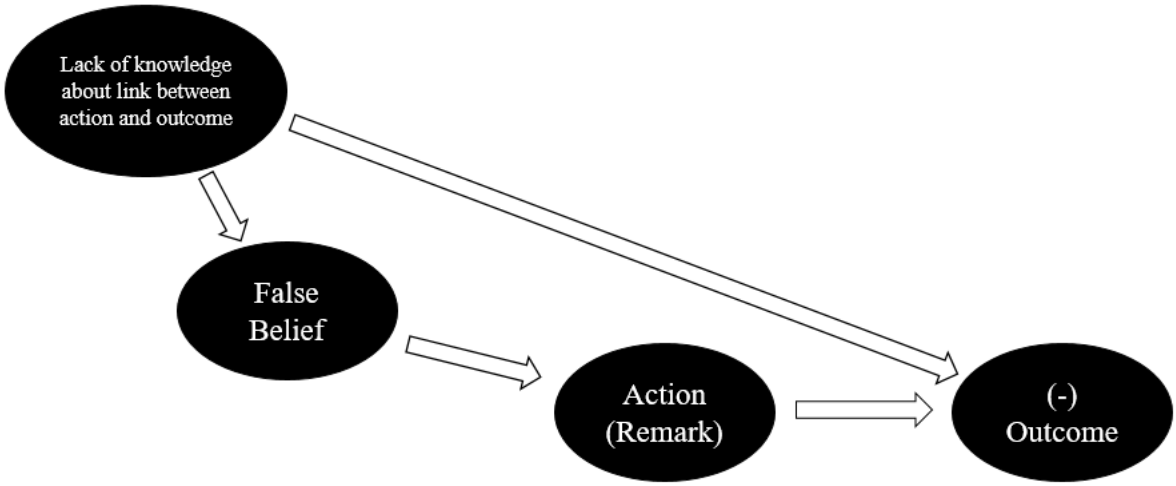

Figure 4.3. In background integration, the agent's lack of knowledge about the link between the action and the outcome can serve as either (1) an explanation for the false belief behind the agent's intention or (2) a direct causal explanation for the unintentionally produced outcome.

Crucially, while the task of integrating mental state information with action, outcome, and background information is conflated with the task of straight-out mental state inference in the No Information condition, this conflation is eliminated when various components of intentional action – specifically, belief and desire – are made explicit. When one or more of these components is made explicit, the observer has the opportunity to draw causal connections between explicit components and to recognize potential inconsistencies among these components, thereby demonstrating an understanding of mind-action integration, mind-outcome integration, and background integration, even in the absence of a capacity to make accurate individual mental state inferences.

In the following section, I outline how an ideal social perceiver would integrate mental state information in each of the information conditions, as well as possible deviations from this ideal ASD adults may exhibit.

4.2.2. Ideal Social Perceiver Analysis

No Information. In the No Information condition, only basic information about the context and the character's utterance is provided – no mental states. In order to demonstrate a full understanding of the action, the observer must (1) *infer* the contents of the agent's belief and positive or neutral desire from the presented background and from general knowledge and (2) successfully engage in the two primary types of information integration. To successfully achieve *mind-action* integration, ASD individuals would have to avoid attributing a negative intention to the speaker (as a subset did in Zalla et al., 2009), an explanation that fails to link the speaker's positive desire to form a correct description of the intentional action. Instead, they would have to see the speaker's positive desire as directly relevant to the action that she carries out. In addition, even if they successfully recognize that the intention *behind* the action was positive, to successfully achieve *mind-outcome* integration, ASD individuals would have to avoid assuming that the positively valenced intention was actually realized, instead seeing that the outcome of the action is (1) offensive to the listener and/or (2) potentially awkward for the speaker. Previous work (Zalla et al., 2009) suggests that, under these conditions, while ASD individuals may effectively infer the agent's general positive desire, they may fail to appreciate that this desire was not actually realized because the agent had a false belief, and may thus fail to understand the negative outcome for the listener (offense) or for the speaker (awkwardness).

Finally, ASD individuals can demonstrate *background* integration by explaining how the agent's lack of knowledge (that her action would lead to the undesirable outcome) served either as a direct causal explanation for the negative outcome, or as a background cause for the agent's false belief.

Explicit desire. In this condition, the observer is presented with the agent's explicit positive desire (e.g., to help the person in the hallway). First, the social perceiver must show a basic recognition that the stated desire is causally linked to the action, and thus the speaker's intention in performing the action is a positive one (*mind-action* integration). Secondly and most centrally, the perceiver must recognize that this desire (e.g., to help someone else) can still lead to an outcome that is not only unintended, but directly conflicts with (and in fact precludes the fulfillment of) the original, desired outcome (e.g., the recipient does not experience the remark as kind; instead, she is insulted by the remark, *mind-outcome* integration). But if the social perceiver fails to recognize that the outcome is either awkward for the speaker (the speaker realizes, after acting, that her earlier belief is or may be false) or offensive to the listener, then he is merely parroting back the explicitly stated mental state information without displaying an understanding of its linkage to the resulting action.

In the presence of an explicit desire, the observer's ability to generate the correct belief demonstrates that he can (1) recognize the puzzle of intentional action created when a single action is brought about by a sincere desire to offer help, but instead results in an insult and then (2) use this conceptual information as a guide to resolving this puzzle by drawing an appropriate inference from other information in the story (e.g., generating a belief: since the agent is clearly just trying to be nice, she must make this unfortunate remark for some other reason: she thinks that James' mother is herself a student).

Explicit false belief. In this condition, the observer is presented with the content of the agent's explicit belief (e.g., about the identity of the person in the hallway). In order to demonstrate *mind-action* integration, the observer must recognize the causal relevance of the

stated belief to a proper description of the agent's intentional action in terms of a plausible desire (e.g., she was just trying to talk to a student in the hallway). To achieve *mind-outcome* integration, the social perceiver must also recognize that the outcome that results is not the outcome that the agent desired to achieve with her action – that is, not the one that would have resulted, had her belief been true instead of false. If the social perceiver demonstrates the basic relevance of the stated belief to the speaker's action (*mind-action* integration), then only one additional piece of information is needed to establish mind-outcome integration. For example, in the presence of mind-action integration, recognition that some negative outcome has occurred as a result of the action (offense of the listener, or 'awkwardness' felt by the speaker) can also signal this understanding, as can a simple recognition that the stated belief is false, because recognition of the belief's falsity also implies that the action cannot have led to its intended outcome.

Explicit desire and explicit false belief. Finally, when both belief and desire are explicitly stated, the observer demonstrates an intact *mind-outcome* integration capacity by recognizing that in spite of a positive desire, the action still leads to a negative outcome. In addition, if the observer is incapable of detecting the puzzle in either of the single-mental state conditions – recognizing the mismatch between an explicit desire alone paired with the action and outcome, or between the explicit false belief alone paired with the action and outcome – then the explicit presentation of both mental states provides the opportunity to demonstrate the ability to do so under "full information" conditions.

4.3 General Methods

4.3.1 Mental State and No Information Conditions

We constructed eight new faux pas story contents. "No information" stories were constructed using the same basic structure of Baron-Cohen et al.'s (1999) original stories, but with several modifications. First, information about the background context was added to each story. Usually (but not always) this information appeared at the beginning of the story. This information was meant to provide a knowledge base for the social perceiver's later interpretation of the utterance. Specifically, it was intended to serve as an affordance for the social perceiver's inference of the character's belief state. Secondly, Baron-Cohen's original faux pas task contained several utterances, sometimes made by the speaker of the focal faux pas utterance. These utterances introduced additional intentions to the story and to the role of the speaker of the faux pas, further complicating the story's structure. We minimized these complications by presenting only the single utterance and its consequences.

Each participant saw 8 unique faux pas story contents, with two stories in each of the four information conditions (No information, Explicit Belief, Explicit Desire, and Explicit Belief & Desire).

4.3.2 Procedure

Participants were instructed that they would read a number of stories "drawn from a larger group of stories varying in difficulty" and to complete the internet-based task entirely on their own. Both ASD and control participants completed Section I with 8 stories. ASD participants had the further option of completing a Section II containing three additional stories (described in detail below). After reading each story on the initial screen, participants answered

a series of questions, each on a separate page. To avoid a high demand on working memory (and to maintain as much consistency as possible with the in-person procedure used in Zalla et al., 2009), each question was presented below the story text.

Foil Stories. In addition to the faux pas stories, we created six "foil stories." In previous versions of the faux pas task (e.g., Baron-Cohen et al., 1999), control stories used very similar contents to faux pas stories, yet all were presented within-subjects. These control stories may have actually misled autistic participants into assuming that the faux pas stories should be interpreted similarly to the control stories, depressing performance. To avoid this problem, we constructed 6 foil stories that described completely different situations from those described in our faux pas stories. In addition, whereas earlier versions of the faux pas task contained only negatively valenced control stories in which the main character had a true rather than a false belief (or lack of knowledge), we introduced a stimulus context in which scenario valence also varied, creating a 2 (Positive or Negative Scenario valence) X 2 (Presence or Absence of False belief) design, where the eight main faux pas stories occupied the "Negative Valence, False belief" cell. Each of the remaining cells contained two foil stories each.

In addition to more precisely defining the nature of autistic deficits with respect to intentional action understanding, the present task also sought to provide greater precision in interpretation of both forced-choice and open-ended questions. To this end, we made a number of modifications to measures used in previous versions of the faux pas task. A complete description of these measures and modifications thereof follows.

Describing the utterance. First, participants were prompted to describe the main character's utterance. Previous studies used a two choice "faux pas detection" question asking "whether or not someone said something wrong/awkward" (Baron-Cohen et al., 1999; Zalla et al., 2009). We reformulated this question to allow participants to describe the utterance using one or more choices. Participants were instructed: "Look back at what [speaking character] said. How would you describe what [s/he] said?" Participants were instructed to "check all that apply" among four options: "It was awkward," "It was nice," "It was mean," and "It was neutral." The order of options was randomized for each story.

Explanation Question. Participants then responded in a text box to the "explanation" question, which simply asked, "Look back at what [character] said. Why did [s/he] say that?" This was reformulated from the original version that appeared in Zalla et al., (2009), "Why was it wrong/awkward?" Although Zalla et al. (2009) classified responses to this question as behavior explanations, this question actually prompted participants to offer explications of the *norm violation*, e.g., why the behavior broke a rule, rather than of the behavior itself. To invite participants to actually explain the cause of the behavior (regardless of their interpretation of the presence or absence of a norm violation), we presented participants with the reformulated question: "Why did he/she say that?" These responses could subsequently be content coded to generate three important variables: (1) Whether participants cited reason explanations – mental states that the explainer took to be on the agent's mind as the reasons for which she acted – or non-mentalistic background factors; (2) the degree to which they demonstrated an understanding that the character's belief state (and its falsity) was highly relevant to explaining the character's

action and/or the resulting negative outcome; and (3) the degree to which the participants incorrectly inferred that the agent acted on the basis of a negative intention.

Forced-Choice Belief Question. Participants then responded to the forced-choice belief question: "Did [character] believe that [contents of the true/false belief]?" Within the foil stories, one of the "false belief" stories was reverse coded, such that the correct answer to half of the false belief stories was "yes" and the other half was "no." One of the faux pas stories was also reverse coded. In addition, our wording of the belief question differed from previous versions of the faux pas task, which explicitly defined the true content for participants as a part of the question, e.g., "Did the character know [true fact from the story]? In adopting this more neutral phrasing, we sought to avoid inadvertently providing additional clues to the question's correct answer.

Empathy Question. The empathy question asked: "How did [recipient] feel after [speaker] said that?" This question was unchanged from its use in Zalla et al. (2009).

Empathy Post-test. Across all participants, there were between 8 and 20 unique responses to the 'empathy' question for each of the 32 faux pas stories (8 story contents in 4 distinct conditions). These responses were used to create a rating task. In this rating task, a new group of typically developing participants first read one of the 32 story versions, and provided their own response to the empathy question. (These open-ended responses were collected to appropriately simulate the original participants' task; the new participants' open-ended responses themselves were not of interest). Then, participants were instructed to read each story again, and

to rate the accuracy of previous participants' responses as an answer to the empathy question on a 1 (not at all accurate) to 7 (very accurate) scale.

A total of 225 typically developing individuals recruited on Amazon Mechanical Turk completed the task online in exchange for monetary compensation (7 per story content in each condition). Participants' mean scores for each unique response were then assigned to the original participants' responses (replacing the text these original participants had provided). The main analyses reported in the results section were performed on these ratings.

Control Questions. Following these questions, participants answered two control questions, which asked about basic details provided in the story (e.g., occupations of the characters or locations). Responses were coded as correct or incorrect by a research assistant.

Translation. Stories were translated into French by a bilingual native French speaker. A French-speaking scientific collaborator reviewed the stories and suggested revisions. Finally, a bilingual native English speaker back-translated the stories. In addition, for purposes of cultural sensitivity, French names were used for the story characters. All open-ended responses (to the control questions, empathy question, and explanation question) were translated from French into English by a native French speaker fluent in English.

Design. Each of the eight faux pas story contents was classified as "short" or "long" according to its word count, and stories were grouped into four pairs, each containing one long story and one short story. The first form was created by assigning one pair of stories to each of

the four main within-subjects conditions, and three additional forms were created by rotating the initial story-condition assignment such that each pair of stories was assigned to each condition once. There were two distinct orders of condition presentation, creating an additional four forms (for a total of 8 unique forms). In both orders, conditions were presented in two blocks of gradually building information, beginning with the short-No information stories, and with the six foil stories interspersed. For example, in the first block, participants first saw a story in the "No information" condition, followed by a story in the "Desire only" condition, followed by a story in the "Belief and Desire information" condition. Each participant saw each story in only a single condition.

Section II: "Teaching" Section. In addition to the conditions in which participants received full mental state information, in the "learning" section, participants again saw three faux pas story contents that they had most recently seen in the No Information condition, the Belief-only condition, and the desire-only condition, only this time, the story appeared in in the full mental state information (belief and desire) condition. While full-information stories yoked to the Desire and No Information conditions were each based on four distinct story contents, Full-information stories yoked to the Belief condition were based on all eight possible story contents. If participants struggled in their performance on a story with less information, they could demonstrate improvement for that same story content in the presence of both mental states.

4.4 Study 6

4.4.1 Methods

Participants. 25 participants (6 Female) with a diagnosis of Autism Spectrum Disorder were recruited from Albert Chenevier Hospital in Creteil, France. All diagnoses were made by experienced clinicians and were based on the DSM-IV TR (American Psychiatric Association, 2000) or DSM-5 (American Psychiatric Association, 2013) and the ASDI (Asperger Syndrome Diagnostic Interview) (Gillberg et al., 2001). Retrospective interviews with parents or caregivers using the ADI-R (Autism Diagnostic Interview) (Lord et al., 2000) or interviews with patients using the ADOS (Lord et al., 2000) confirmed the diagnoses. 38 typically developing control participants were recruited from a sample of individuals whose age, gender, and IQ profiles were similarly distributed to the overall population of ASD participants. All participants completed the main experimental task online.

For all participants, IQ was measured using the Wechsler Adult Intelligence Scale (WAIS, fourth edition; Wechsler, 2008). To achieve a desired statistical power of .8 and detect an effect size of .86 (the smallest reported effect size in Zalla et al., 2009), the desired final *N* for ASD participants is 30, and data collection is still ongoing. Once IQ scores for a sufficient number of control participants have been collected, a subset of the controls will be selected to match the ASD group as closely as possible on age, sex, and IQ. The present analyses report data from all available control participants.

Clinical characteristics of the sample. Existing data on the sample are provided in Table 4.1. For control participants, IQ scores reflect the mean of 17 available scores (out of 38 total controls), while age statistics were available for 32 control participants. Gender data were available for 36 control participants. Two ASD participants were missing IQ data.

Participants completed the main task online in exchange for monetary compensation or an Amazon gift card.

Table 4.1.

Current demographic information (means listed for Age and IQ, standard deviations in parenthesis)

	Gender	Age	IQ
ASD	72% Men (*N* = 18)	34.56	110.76
		(10.05)	(13.48)
Control	50% Men (*N* = 18)	27.4	111.74
		(7.12)	(17.88)

4.4.2 Results and Discussion

Participants' responses on a range of forced-choice and open-ended questions provided insight into both their abilities to infer mental states from complex behavioral information, and their abilities to integrate explicitly presented information into a correct interpretation of intentional action and its consequences in the world. Results on each of the forced-choice measures are reported first, followed by results from the explanation and empathy questions. Each results section is followed by a brief discussion. In general, within-subjects comparisons of interest are framed as three central ANOVA contrasts: (1) the No Information condition compared with the average of the three conditions in which explicit mental state information was provided (a test of the "inference" hypothesis"), (2) the Explicit Belief condition compared with the Explicit Desire condition, and (3) the Explicit Belief (alone) and Explicit Desire (alone)

conditions compared with the condition in which both Belief and Desire were presented explicitly (all tests of the integration hypothesis).

Control Questions. Participants received a "1" for a correct response and a "0" for an incorrect response for each of the two control questions asked per story. Scores were then aggregated by condition and averaged, such that each participant's score reflected the average number correct per story (out of a total of two questions per story). There was no overall difference in the mean number of control questions answered correctly between ASD and control participants [$M_{Control}$= 1.98, SD = .06; M_{ASD} = 1.96, SD = .08, $t(58)$ = .65, ns]. Furthermore, there was no interaction between group (ASD vs. Control) and Information condition ($Fs_{Interaction}$ < 1.13, ps > .29), nor were there any main effects of Information condition (Fs < 2.20, ps > .14).

Interpretation of the utterance. For each utterance, participants had the option to check "nice," "mean," "awkward," and/or "neutral." Responses for each of the four variables were first aggregated across each cell of the 2 (autism vs. control) X 4 (mental state information) design. There was a high degree of similarity of correlation matrices across all 8 cells of the design, and this was confirmed by Box's M test for homogeneity of variance-covariance matrices, M = 193.24, $F(136, 8292)$ = .98, p = .54. Values on each of the descriptor variables were aggregated across participant group and information condition, yielding one variable for each descriptor with a score ranging from 0 to 8 (indicating the number of stories, out of 8 total, for which that descriptor had been endorsed), and subsequently entered into a principle components analysis. Two orthogonal components were extracted. The "awkward" and "neutral" variables loaded most highly on the first component, while the "nice" and "mean" variables loaded most highly

on the second component. Component loadings for each of the four variables are provided in Table 4.2.

Table 4.2

Standardized component loadings for the orthogonal 2-component solution for "description of the utterance"

	Component 1	Component 2
Awkward	.98	-.02
Nice	-.30	.71
Mean	-.01	-.74
Neutral	-.59	.44

Theoretically, the two components capture the intention underlying the action (nice or neutral vs. mean, Component 2) and the outcome of the action (nice or neutral vs. awkward, Component 1). Component scores for each participant were computed from linear combinations of the four constituent variables, which ranged in value from 0 to 2 (indicating the number of stories—out of 2 possible per information condition—for which the participant had checked each descriptor). Possible component values thus ranged from approximately -2 to 2, with (for example) a score close to 2 on Component 1 (outcome) indicating the participant had checked "awkward" for both stories in that condition but no other options, and a score close to -2 indicating that "nice" and "neutral" were checked for both stories, but not "awkward." Because the two components were derived with an orthogonal rotation and thus uncorrelated (r = .04, *ns)*, control vs. autistic participants' scores on each component were compared across the four conditions using two separate (univariate) mixed ANOVAs. Means and standard deviations for both components are depicted in Figure 4.4 (outcome) and Figure 4.5 (intention).

There was no significant main effect of information condition for either the intention or the outcome component (all *F*s (1, 59) < 1.97, *ns*), nor were there any interactions for the intention component (all *Fs* < 1.13, *ns*). Control participants endorsed "awkward" versus other outcome descriptors with marginally greater frequency than ASD participants in the No Information condition $t(61) = 1.74$, $p = .09$, and they also endorsed it more ($M = 1.36$, $SD = 0.63$) than they did in the three mental state information conditions, while ASD participants' performance between the no information and mental state conditions remained relatively constant, $F_{interaction}(1, 59) = 5.21$, $p = .03$. In the presence of explicit mental state information, there was no difference between ASD and control participants.

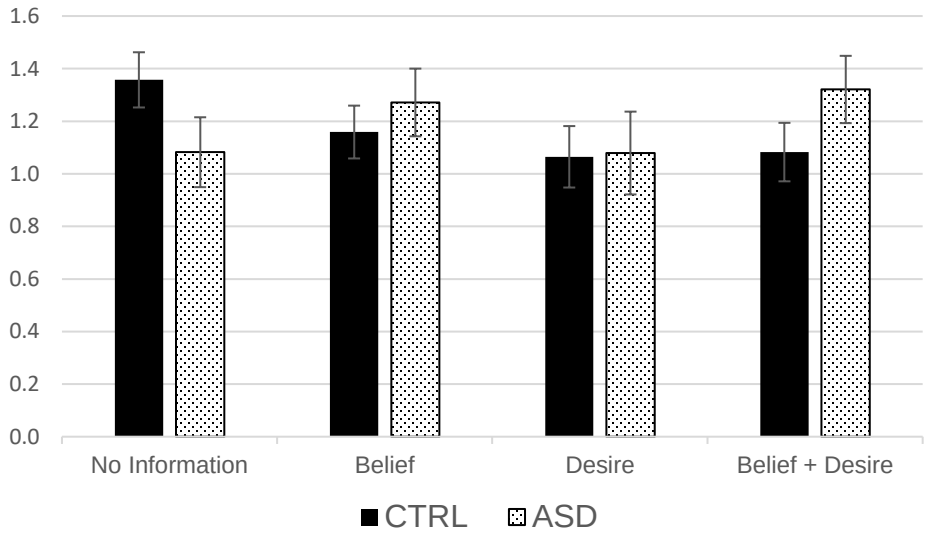

Figure 4.4. Scores on the "outcome" component (where scores closer to +2 indicate endorsement of "awkward" and those closer to -2 indicate "nice" or "neutral.)"

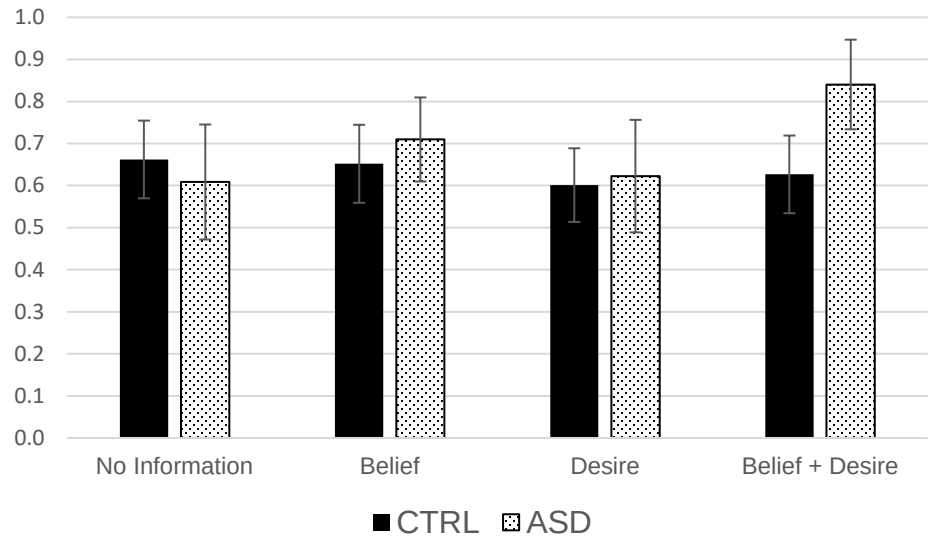

Figure 4.5. Scores on the "Intention" component (where scores closer to 2 indicate "nice" or "neutral" and scores closer to -2 indicate "mean.")

Teaching Section. In the teaching section, ASD participants each received three stories they had previously seen in either the No Information, Belief Only, or Desire Only conditions in the Full Information (Belief + Desire) condition. Difference scores for the Outcome and Intention components were first calculated by subtracting participants' scores on original stories in the Belief only, Desire only, or No Information conditions from those same story contents presented again in Part II in the Belief + Desire condition. (These component scores are reported in Table 4.3.) A positive difference score on the Outcome component would indicate a greater tendency to endorse "awkward" versus "nice" or "neutral" in the presence of full mental state information than with incomplete mental state information, and a positive difference score on the Intention component would indicate a greater tendency to endorse the "nice" versus the "mean" or "neutral" descriptors when presented with full mental state information versus incomplete (no information, belief only, or desire only) information. A one-way within-subject ANOVA compared differences from each of these three information conditions with the full information

condition (Belief + Desire condition - Original story). For the outcome component, a planned contrast revealed no significant difference between adding both mental states to an original "No Information" story, and adding an additional mental state to a story that already contained a mental state, $F(1, 22) = 0.57$, $p = .46$. However, when a story content was presented in the Belief + Desire condition in section II after that same story content was presented in the Desire-only condition in Section I, ASD participants rated the outcome of the full-information story as *less* awkward (Outcome component = 0.46) than the original desire-containing story (Outcome component = 0.58), and this differed from the slight *increase* in endorsement of awkwardness seen in response to Full Information stories that had previously been presented in the Belief-only condition, $F(1, 22) = 7.36$, $p = .01$.

For the intention component, when a story content was presented in the Belief + Desire condition in Section II after being presented in the No Information condition in Section I, ASD participants increased their endorsement of the "Nice" descriptor more than they did when the Belief + Desire story was preceded by a story of the same content in one of the single mental state conditions (belief or desire only), $F(1, 22) = 5.32$, $p = .03$. In addition, the increase in endorsement of the "Nice" descriptor was greater in response to Belief +Desire stories that originally appeared in the Belief only condition than it was from those that had originally appeared in the Desire only condition, $F(1, 22) = 3.71$, $p = .07$.

Table 4.3. *Means (standard deviations) of component scores for outcome and intention in the "Teaching" section (presented in the full information, or "Belief + Desire" conditions) and the conditions in which those same story contents appeared in Section I (in the No Information, Belief alone, or Desire alone conditions). Only ASD participants completed the teaching section.*

	No Info	Belief + Desire (NI)	Desire	Belief + Desire (D)	Belief	Belief + Desire (B)
Outcome	0.48	0.44	0.58	0.46	0.27	0.50
	(0.50)	(0.50)	(0.45)	(0.50)	(0.42)	(0.48)
Intention	0.24	0.25	0.45	0.56	0.05	0.42
	(0.48)	(0.50)	(0.46)	(0.36)	(0.35)	(0.46)

Note. Because component scores are derived from single stories rather than aggregates of two stories (as in the main analysis), scores range from -1 to 1 rather than -2 to 2. "Belief + Desire (NI/B/D)" refers to stories in the teaching section that were previously presented in the No Information condition, Belief only condition, or Desire only condition, respectively.

Interpretation of the Utterance: Discussion. In the discussion section for this and every subsequent measure, I first discuss evidence for the two competing accounts of inference from No Information condition and then continue on to the discussion of integration competence in the three explicit information conditions.

In the No Information condition, ASD participants tended to describe the speaker's intention as nice or neutral with the same likelihood as control participants, and both groups tended to describe the intention more as nice or neutral than as mean. On the "outcome" component, ASD participants endorsed the "awkward" descriptor marginally less frequently (and the "nice" or "neutral" descriptors more frequently) than did control participants in the No Information condition. While control participants saw the No Information scenarios as highly

awkward, this assessment decreased in the presence of explicit mental state information. Perhaps because awkwardness would generally result only when the speaker became *aware* of the insulting nature of his remark, the emphasis on the speaker's mental states prior to having made the remark may have shifted the assessment toward the "nice" or "neutral" descriptors, eliciting a "curse of knowledge" of sorts (Birch & Bloom, 2004): knowledge of the speaker's mental state *prior* to performing the action may have made it more difficult to grasp that same speaker's changed mental state *after* performing the action. In contrast, ASD adults were not subject to such a curse: their detection of awkwardness in the No Information condition did not differ from their own or TD participants' detection of it in the explicit information conditions. This pattern of results suggests that control participants responded to the presentation of explicit information differently than ASD individuals, but does not definitively suggest that ASD individuals suffered any particular deficit in this condition either, especially since ASD participants' performance in the No Information condition was similar to that of control participants in the three explicit mental state information conditions. Thus, there is no definitive evidence for either account: while we cannot rule out that the marginal difference between ASD and control participants in the No Information condition is due to an enduring mental state inference deficit, the fact that this performance is both only marginally different from controls' and is no worse than that of controls in the other three conditions, suggests that ASD individuals are performing at a similar level to control participants overall. Thus, the knowledge-based inferential deficit cannot be rejected either.

However, there was clearer evidence that ASD and control participants performed comparably in the three explicit mental state information conditions on the outcome component, suggesting that ASD participants are capable of using belief and desire information to understand

the valence of a character's intention (in the service of successful *mind-action* integration**)**, as well as to detect (at least as well as controls under similar conditions) when the speaker has said something that he may later realize would have an unintended effect (in the service of successful *mind-outcome* integration). In addition, in the Teaching section, ASD participants were also responsive to the addition of "full" (both belief and desire) mental state information to stories that originally contained only a single mental state. In particular, ASD participants decreased their endorsement of the 'awkward' descriptor when a belief was added to a story that had previously had only explicit desire information, while they increased it when a desire was added to a story that had previously had only explicit belief information. While presenting the speaker's false belief may have distanced the speaker from any awareness of the negative outcome in the eyes of ASD participants (perhaps leading them to conclude that he did not know what he was saying, and would not realize its impact after saying it), presenting desire information may have more starkly highlighted the direct conflict between the speaker's desired outcome and the actual outcome. In other words, the addition of desire information may have increased ASD participants' awareness of the one of the action's central puzzles: the fact that the speaker has a positive or neutral desire that goes unfulfilled in the story. They demonstrated this awareness by more frequently endorsing the 'awkward' descriptor. Results from the teaching section suggest, then, that the addition of desire information was more helpful in facilitating **mind-outcome** integration for ASD participants than was belief information.

Forced Choice Belief question.

***Data preparation*.** Two control participants were excluded from the analyses due to missing data in two cells.

Correct responses to the belief question were aggregated across the two stories within each of the four information conditions, yielding a score of 0 to 2 per condition. Means for each group across the four conditions are reported in Table 4.4.

Table 4.4

Mean (standard deviation) of number correct (out of 2) on the belief question

	No Information	Belief	Desire	Belief & Desire
ASD	1.64	1.80	1.88	1.84
	(0.57)	(0.41)	(0.33)	(0.37)
Control	1.81	1.89	1.67	1.83
	(0.40)	(0.32)	(0.48)	(0.38)

To examine the main effect of adding additional information on responses to the belief question, a 4 (Information condition) X 2 (Autism vs. Control) mixed ANOVA was conducted. Three planned contrasts examined main effects for the information factor across all participants. Compared with the no information condition, conditions in which explicit mental state information was provided (the Belief, Desire, and Belief + Desire conditions) provided a marginal improvement in performance that did not reach significance (F (1, 60) = 2.26, p = .11). There was no significant main effect of receiving both mental states in comparison to either belief or desire alone, F(1, 60) = .20, p = .65) nor was there a main effect for receiving belief versus desire information, [F (1,60) = 1.18, p = .28)].

Interactions of each of these within-subject contrasts with the autism factor were then examined. The interaction of the autism factor with the Information vs. No information contrast

approached, but did not reach, significance, $F(1,60) = 3.23$, $p = .08$, suggesting that over and above general improvement for both groups in the presence of mental state information, the autism group reaped some additional benefit. However, ASD ($M = 1.64$, $SD = .57$) and control ($M = 1.82$, $SD = .39$) participants performed similarly in the no information condition (Welch's $t(38.95) = 1.35$, $p = .185$). Unlike controls, ASD participants also benefited from the presentation of explicit desire information; controls benefited only in the presence of belief information [$F(1, 60) = 5.57$, $p = .02$].

Part II: "Teaching" section. Similar to scores for the Outcome and Intention components, difference scores for the Teaching Section for the forced-choice belief question were first calculated by subtracting ASD participants' component scores on original stories in the Belief only, Desire only, or No Information conditions from the component scores for those same story contents presented again in Part II in the Teaching Section (where all stories appeared in the Belief + Desire condition). Two participants who did not complete the teaching section were excluded from analysis. A one-way within-subject ANOVA compared improvement from each of these three information conditions to the full information condition. Two planned contrasts revealed no significant performance difference between adding both mental states to an original "No Information" story, and adding an additional mental state to a story that already contained a mental state, $F(1, 22) = 0.72$, $p = .41$. There was also no difference between adding a belief to a story that originally contained a desire, and adding a desire to a story that originally contained a belief, $F(1, 22) = 2.10$, $p = .16$. In summary, receiving previously read stories in the Belief + Desire condition after having viewed them under less favorable information conditions did not improve performance on the forced-choice belief question.

Forced-Choice Belief Question: Discussion. In response to the forced-choice belief question, ASD participants performed similarly to typically developing controls in the No Information condition. This finding suggests that the inclusion of additional background information in this set of faux-pas scenarios (compared with previous ones) successfully facilitated belief inference for ASD individuals; in turn, this finding provides support for the knowledge-based inferential deficit account. While autistic participants' rates of correct belief inference benefited from receiving both belief and desire information (presented both individually and together), control participants benefited only in the presence of an explicit belief (either alone or presented with a desire). It is unsurprising that performance on the belief question would improve in the presence of explicit belief information; in this condition, this question serves as a manipulation check. However, it is notable that ASD participants performed at an equally high level when only a desire was provided, suggesting that they were able to successfully engage in *mind-outcome* integration. To this end, they combined information about the character's desire with other information presented in the story to detect the overall intentional action puzzle and resolve it with the missing belief: although the character had a benign desire, her action resulted in a negative outcome, and the missing belief accounted for this.

However, the performance of control participants was curious; while they benefited from belief information, their correct identification of the belief declined in the presence of explicit desire information alone. Because mentioning the character's naïve desire in the story was superfluous from a pragmatic perspective (assuming the "No Information" version of the story successfully affords an inference of the character's desire without it being explicitly mentioned),

control participants may have engaged in excessive inferencing in this specific condition. For example, they may have inferred that the presence of the desire actually meant that in *spite* of having a true belief and making an apparently mean comment in light of that true belief, the person's intention truly was a benign or a positive one. They may have reasoned the speaker was a generally naïve person. In contrast, autistic participants, who have documented deficits in pragmatic communication (e.g. Klin & Volkmar, 1997; Tager-Flusberg & Sullivan, 1995), may not have recognized the pragmatic violation of providing the character's desire in the story itself, and simply used the information to enrich their existing understanding of the story.

Responses to the Explanation question: Conceptual Form

Coding scheme. The conceptual form of responses to the "why" question was determined using the Folk Explanations (F.Ex) coding scheme (Malle, 1998, 2004) for open-ended behavior explanations. This scheme categorizes explanations of intentional actions into reason explanations and causal history of reason (CHR) explanations. Within reason explanations, it distinguishes between belief and desire reasons, as well as between reasons expressed using mental state markers (e.g., "He knew, he hoped") and reasons cited without such markers. Within causal history explanations, it distinguishes between stable traits (dispositional properties such as personality and character) and non-traits.

F.Ex Coding Reliability. In Study 6 there were a total of 14 stories each for 39 control participants, and 17 stories each for 21 ASD participants (with the exception of two ASD participants who elected not to complete the teaching section). For reliability purposes, two coders classified explanation responses to 129 items from Study 6. Coders reached 91%

agreement on the codability of explanations (κ = .35)[10], 98% agreement on distinguishing reason explanations from causal history explanations (κ = .79), 87% agreement on distinguishing belief reasons from desire reasons (κ = .87), and 95% agreement on the application or non-application of mental state markers (e.g., "he thought, she wanted") to reason explanations, (κ = .62). In addition, agreement on classification of causal history explanations appealing to some feature of the person (e.g., the person's behavior, internal state, attention, or perception) was 85% (κ = .87) and agreement on classification of trait vs. nontrait causal history explanations was 100% (κ = 1.0). All remaining items were classified according to the F.Ex coding scheme (Malle, 1998/2010) by a single coder (involved in the above process of reaching reliability) blind to participant group and hypotheses.

Results. All means reflect the number of explanations given *per story*, aggregated across the two stories within each of the four information conditions. For example, if a participant gives 1 reason and 2 causal histories for story 1 in the No Information condition and 2 reasons and 1 causal history for story 2 in that condition, that participant has a mean of 1.5 reasons and 1.5 causal histories per story for that condition. Means and SDs for Reason vs. CHR explanations are displayed in Figure 4.5, and means for these as well as Belief vs. Desire reasons and Marked beliefs versus Unmarked beliefs, are reported in Table 4.5. Mixed (Information X Autism) ANOVAs were performed on three sets of theoretically relevant difference scores: Reasons - CHRs, Beliefs - Desires, and Unmarked beliefs - Marked beliefs (participants provided too few markers for a similar analysis to be performed for desires). In addition, a

[10] Although reliability was calculated separately for Study 6 and Study 7, the two coders coded parts of Study 6 and 7 together for reliability training. Study 6 explanations were only the first step in the training; κ for codability was improved (see reliability, Study 7) before the second coder coded either study independently.

planned analysis of Trait vs. Non-trait CHRs could not be performed due to too few ASD participants who provided trait explanations.

Reasons and Causal Histories. Control and ASD participants gave comparable numbers of reason and causal history explanations for stories that did and did not contain explicit mental state information $F(1, 56)_{\text{Autism X Mental State Information interaction}} = .74$, $p = .39$), as well as comparable numbers of these explanation types in the presence of a single versus both mental states, $F(1, 56)_{\text{interaction}} = 0.01$, $p = 0.92$). There was, however, a significant interaction between information and autism for the comparison of the explicit belief versus explicit desire conditions, $F(1, 56) = 4.43$, $p = .04$), with ASD participants providing fewer CHRs and more reasons in the belief condition than in the desire condition, while control participants provided a relatively constant rate of CHRs and reasons across the two conditions.

Reasons: Belief and Desire Reasons. ASD and control participants gave comparable numbers of beliefs and desires across the four information conditions, $Fs_{\text{interaction}} < .70$; $ps > .40$).

Beliefs: Mental State Markers. Control participants gave marginally more marked than unmarked beliefs in the No Information condition than did ASD participants, $F(1, 34) = 2.87$, $p = 0.10$. However, this difference was not driven primarily by differential use of marked beliefs, as the two groups gave comparable numbers of marked beliefs in the No Information condition, $t(43) = .518$, $p = .61$. ASD and control participants gave comparable numbers of marked and unmarked beliefs across the other three information conditions, $Fs_{\text{interaction}} < .08$, $ps > .78$.

Causal histories: Traits. Too few participants gave trait explanations for formal analyses (as few as 4 participants per cell for difference scores). The raw number of traits for all participants was small in all conditions for both groups (e.g., for No Information, $M_{Autism} = 0.16$, $SD = 0.38$; $M_{Control} = 0.12$, $SD = 0.39$; for Explicit Belief, $M_{Autism} = 0.00$, $SD = 0.00$; $M_{Control} = 0.06$, $SD = 0.22$).

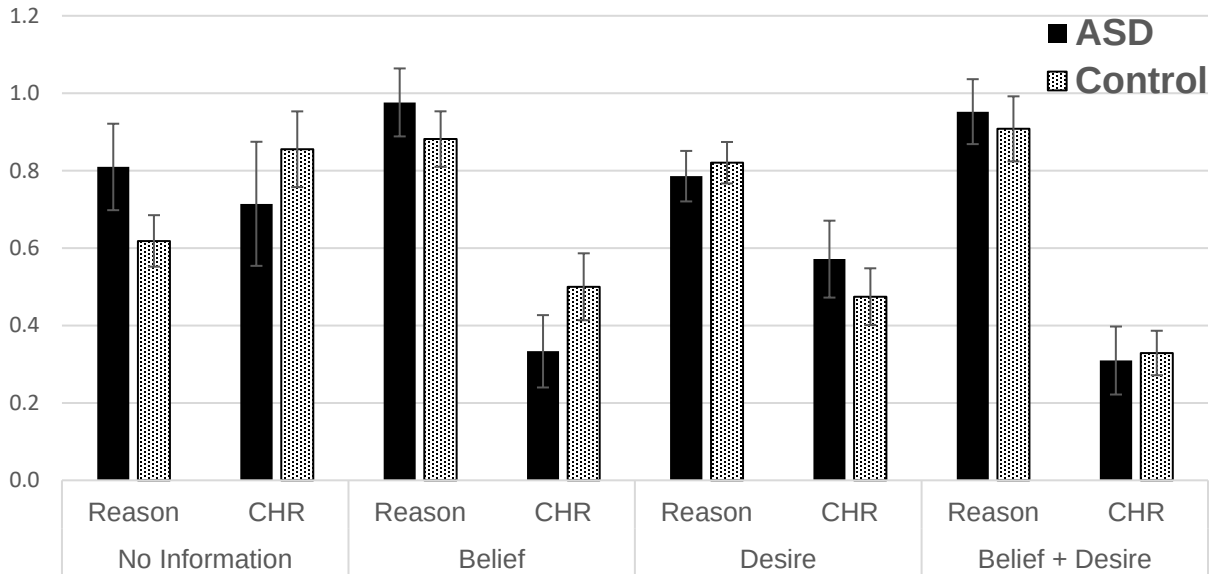

Figure 4.5. Number of Reason and CHR explanations (per behavior) given in each of the four information conditions.

Table 4.5.

Mean (standard deviation) number of explanations per behavior in each of the four information conditions. Above: Comparisons of Reasons vs. CHRs. Middle: Comparison of Belief and Desire reasons. Below: Comparison of Marked and unmarked beliefs.

	No Info		Belief		Desire		Belief + Desire	
	Reason	CHR	Reason	CHR	Reason	CHR	Reason	CHR
ASD	0.81	0.71	0.98	0.33	0.79	0.57	0.95	0.31
	(0.52)	(0.73)	(0.40)	(0.43)	(0.30)	(0.46)	(0.38)	(0.40)
Control	0.61	0.86	0.88	0.50	0.82	0.47	0.91	0.33
	(0.41)	(0.60)	(0.44)	(0.53)	(0.33)	(0.46)	(0.52)	(0.35)
	Belief	Desire	Belief	Desire	Belief	Desire	Belief	Desire
ASD	0.94	0.21	0.91	0.21	0.75	0.35	0.88	0.21
	(0.35)	(0.36)	(0.26)	(0.41)	(0.47)	(0.40)	(0.42)	(0.34)
Control	0.85	0.21	0.86	0.29	0.73	0.36	0.93	0.19
	(0.35)	(0.44)	(0.49)	(0.44)	(0.46)	(0.45)	(0.40)	(0.32)
	Marked	Unmarked	Marked	Unmarked	Marked	Unmarked	Marked	Unmarked
Control	0.93	0.09	0.98	0.12	0.91	0.13	0.95	0.13
	(0.29)	(0.27)	(0.36)	(0.41)	(0.34)	(0.30)	(0.42)	(0.34)
ASD	0.88	0.18	0.90	0.12	1.0	0.06	1.03	0.05
	(0.28)	(0.35)	(0.30)	(0.27)	(0.35)	(0.24)	(0.11)	(0.16)

Teaching section. Mean number of reasons, causal histories, beliefs, desires, and each type of marked reason explanation were calculated for Belief + Desire stories that followed the same story contents in each of the incomplete information conditions. Because each participant responded to only one Belief + Desire story for each of the incomplete information conditions

and participants tended to give only a single explanation type, the *N* of valid difference scores was too small ($N = 8$) for any explanation parameter to perform formal statistical analyses.

Quality of Explanation Coding

Quality coding of responses to the explanation question was performed on all faux pas stories. Quality standards were developed by creating answers that an "ideal social perceiver" would give for each story in response to the question, "Look back at what he/she said. Why did he/she say that?" Each "ideal social perceiver" description cited a background fact from the story that may have played a part in producing the false belief (e.g., "Sandeep is used to having his gas pumped for him, since he is from New Jersey,") as well as the content of the person's false belief (e.g., "Sandeep thought the man was a gas station attendant"), the content of the knowledge that the person lacked (e.g., "Sandeep didn't know that the man was just another customer,") and the content of the appropriate, positive or neutral desire (e.g., "Sandeep was just trying to get his gas pumped.") Because participants provided a mean of only 1.34 explanations per story, use of the "ideal social perceiver" standard for participants (which involved providing four distinct explanations) was not practical, but identification of these elements – and especially, the correct content of the false belief/lack of knowledge – provided a guideline.

Two coders, blind to participant group, used these standards to classify each response into a single numbered category, 0-3. To receive a perfect score of 3, the participant had to give an explanation that directly stated or otherwise implied that the story character had a false belief (in the story in which Sandeep mistakenly asks another customer, an older man, to pump gas for him, "Sandeep thought that the man was a gas station attendant," or, "Sandeep said that because he was trying to get the attention of a person who seemed like the gas station attendant") or that

the story character lacked a crucial piece of knowledge ("Sandeep didn't know that the old man was just another customer"). To receive a "2" rating, a response had to suggest or point toward the content of the correct false belief without stating it outright (e.g., stating that Sandeep wanted someone to help him to pump his gas, but no one came). A "1" response cited information from the story that may have been true, and was possibly broadly relevant to false belief understanding, but did not directly state, nor did it necessarily imply, the false belief (e.g., simply restating that Sandeep is from New Jersey). Irrelevant but correct information drawn from the story and inferences about the personality of the speaker that did not incorrectly imply a negative intention also received a "1" response. A "0" response indicated that the participant had mistakenly attributed a true belief or negative intention to the speaker (including incorrect descriptions of the speaker's statement as joking or ironic).

After a first round of initial independent coding based on these guidelines, coders resolved disagreements and developed basic conventions where necessary. The coders then returned to (once again independently) recode the entire dataset based on the agreed-upon framework. Remaining disagreements were resolved in discussion. Where agreement could not be reached, items were excluded from analysis.

Reliability. Two coders classified 532 responses. Overall agreement on all explanation types was 94% ($\kappa = .85$). Consensus on 5 items (0.9% of all items) could not be reached and these items were excluded from analysis.

Quality coding analysis.

"Don't know" responses. The number of "Don't know" responses was negligible, with ASD participants giving a total of 1 (0.1%) such responses and Control participants giving a total of 7 (1.3%).

To examine the main effect of information on the quality of explanation responses, a 4 (Information condition) X 2 (Autism vs. Control) mixed ANOVA was conducted. Three planned contrasts examined main effects for the information factor. Compared with the No Information condition, there was no additional overall improvement from conditions in which explicit mental state information was provided (the Belief, Desire, and Belief + Desire conditions) over performance in the No Information condition, $F(1, 57) = 1.95$, $p = .17$. However, in the presence of both mental states, participants provided higher-quality responses than they did in the presence of a single mental state, $F(1, 57) = 4.34$, $p = .04$. There was no difference in overall response quality between the belief and desire conditions, $F(1, 57) = .02$, $p = .90$). In contrast to the forced-choice belief question, there were no significant interactions between autism and information condition (all $Fs < 0.60$, $ps > .45$). Consistent with the forced-choice belief question, there was no difference between control ($M = 2.17$, $SD = .63$) and ASD ($M = 2.02$, $SD = .66$) participants in the No Information condition, $t(57) = 0.85$, $p = .40$. Means are reported in Table 4.6.

Table 4.6

Mean (standard deviation) of quality score on the explanation question

	No Information	Belief	Desire	Belief & Desire
Autism	2.02	2.21	2.14	2.31
	(0.66)	(0.66)	(0.69)	(0.70)
Control	2.17	2.13	2.17	2.39
	(0.63)	(0.77)	(0.66)	(0.81)

Negative Intentions. All responses were also classified according to whether they explicitly cited an incorrect negative intention of the speaker (e.g., "He was trying to offend her.") To this end, two coders revisited the codes from the initial coding run. Only original codes of "0" were eligible for consideration as citing incorrect negative intentions, because they indicated a complete lack of understanding of the false belief. The coders classified these responses as either citing or not citing an incorrect negative intention, with an agreement of 100% ($\kappa = .83$).

Rates of negative intentions in both the control and ASD participants were near zero in all four conditions (see Table 4.7). Because there was no variance in the control group for two of the four conditions, and no variance in the ASD group in a single condition, significance testing was not performed.

Table 4.7

Mean (standard deviation) number of negative intentions given (out of two possible in each condition)

	No Information	Belief	Desire	Belief & Desire
Autism	0.05	0.05	0.05	0.00
	(0.22)	(0.22)	(0.22)	(0.0)
Control	0.00	0.03	0.00	0.03
	(0.0)	(0.16)	(0.0)	(0.16)

Lack of knowledge. In ratings of explanatory quality, citing the specific contents of the agent's false belief (which explain why the agent *chose* to perform the action she did) was rated as equally correct to specifically citing the circumstances of which the agent was *not* aware. We thus further broke down all responses that received a perfect score of "3" as either mental state contents (reason explanations for an intentional behavior) or "lack of knowledge" contents (serving as a background explanation, or "causal history" of the mental state, as well as a causal explanation for the unintended outcome). A single coder familiar with the structure of the responses performed these classifications. Mean counts for correct (perfect score) belief contents and correct lack of knowledge contents are reported in Table 4.8**.**

For correct lack of knowledge responses, there were no significant differences across information condition due to group membership ($Fs < 0.51$, $ps > .48$). However, there was a main effect of information condition, such that all participants gave fewer correct lack of knowledge responses in the presence of explicit mental state information than in the absence of such information, $F(1, 57) = 4.64$, $p = .04$. For correct belief responses, there were again no significant group by condition interactions, but participants gave more correct beliefs in the

presence of explicit mental state information than in its absence, $F(1, 57) = 3.59$, $p = .06$, more correct beliefs when provided with explicit belief than desire information, $F = .006$, $p = .02$, and more correct beliefs when provided with both mental states rather than just a single mental state, $F = 5.69$, $p = .02$.

Table 4.8. Means (standard deviation) number of beliefs and lack of knowledge ("LOK") responses given in the "3" category, per condition (out of two stories).

	No Info		Belief		Desire		Belief + Desire	
	Belief	LOK	Belief	LOK	Belief	LOK	Belief	LOK
Control	0.71	0.24	0.92	0.08	0.74	0.08	1.13	0.08
	(0.61)	(0.43)	(0.78)	(0.27)	(0.76)	(0.27)	(0.91)	(0.27)
ASD	0.86	0.24	1.10	0.14	0.67	0.19	1.14	0.10
	(0.76)	(0.24)	(0.70)	(0.36)	(0.66)	(0.40)	(0.73)	(0.30)

Explanations for the Behavior: Discussion. Responses to the explanation question ("Why did he say that?") were analyzed both for their conceptual form and for the quality of their content. While they used mental state markers somewhat differently in the No Information condition, ASD and control participants did not differ in the number of marked beliefs they provided. In addition, ASD and control participants gave comparable numbers of mental state (reason) explanations relative to causal background explanations in the presence vs. absence of explicit mental state information, providing more support for the notion that previous struggles on this task can be accounted for by a knowledge-based inference deficit. However, ASD and control participants responded somewhat differently to the presence of explicit belief vs. explicit

desire information, suggesting that ASD individuals on the autism spectrum may take explicitly presented belief information as a particularly strong invitation to provide more mental states. This difference in the use of concepts – toward a greater use of mental states by autistic individuals – suggests a learned sensitivity to the importance of belief information for behavior explanation.

Distinct from classifications for conceptual form, ratings of explanation quality were designed to capture the extent to which participants' responses correctly linked (directly or indirectly) the main story character's mental state to the action or its outcome. Although this was usually achieved by directly citing the character's belief and acknowledging (or at least implying) its falsity, it could also be done by pointing out facts of which the speaker was unaware, or by citing a desire that directly implied the false belief. Notably, even though ASD participants used slightly fewer mental state terms in the No Information condition, this did not translate into a difference in explanation quality between TD and ASD participants in the No Information condition. This suggests that, at least in adults, previous struggles on comparable tasks (e.g., the faux pas and strange stories tasks) can be accounted for largely by a knowledge-based inference deficit. In addition, although the quality of all participants' responses improved in response to the explicit presentation of both mental states, there was no difference in ASD and control participants' improvement in explanation quality from the No Information to the average of the belief, desire, or full information conditions. Comparable performance of the autism group to controls in the desire condition suggests intact *mind-outcome integration*: though neither group improved notably in the Desire condition above performance in the No Information condition, ASD individuals were able to integrate explicitly presented desire information with information about its lack of fulfillment to generate a correct belief to the same degree as

controls. In addition, the additional benefit conferred by the presence of both mental states over that of a single mental state suggests an important role for the integration of desire information for mind-outcome integration. ASD participants also exhibited uniformly good performance on *mind-action* integration, as they provided almost no negative intentions. Finally, ASD and control participants employed *background* integration to roughly the same degree, providing "lack of knowledge" responses at comparable rates in each of the four conditions.

Strikingly, although there were some differences in the rates at which ASD participants appealed to mental states and non-mentalistic explanations and mental state verbs across the four conditions, it does not appear that ASD and TD participants reached their *correct* interpretations by appeal to distinct concepts. In their correct responses, both groups appealed most often to the speaker's lack of knowledge in the No Information condition, and less often in each of the three explicit mental state information conditions. They appealed least to beliefs in the No information condition, and increasingly to beliefs in the presence of an explicit belief or in the presence of both a belief and a desire rather than a belief or a desire alone.

Empathy Question. Analyses for the empathy question were conducted on the post-tested mean accuracy ratings of ASD and control participants' responses provided by a second group of typically developing participants. As in previous measures, scores were aggregated over story within condition, such that the mean score for each condition reflected the mean of the two stories within that condition. There were no differences between the two groups in the No Information condition [$t(53) = 0.47$, *ns*), and no differences between groups across the four conditions ($Fs_{Interaction} < 2.51$, $ps > .12$), but the scores of all participants improved from the No Information condition to the three explicit information conditions [$F(1, 50) = 8.88$, $p < .01$], and

improvement in the Belief condition was higher than improvement in the Desire condition [$F(1, 50) = 7.92$, $p < .01$). Means are reported in table 4.9.

Table 4.9.

Means (standard deviations) for post-test ratings on the empathy question.

	No Information	Belief	Desire	Belief & Desire
Autism	5.31	5.74	5.55	5.83
	(0.73)	(1.17)	(1.07)	(0.58)
Control	5.40	5.96	5.41	6.12
	(0.70)	(0.67)	(0.79)	(1.04)

Regression of main variables on explanatory quality. Ratings of explanation quality, which were developed from the responses of an idealized social perceiver, serve as a useful summary measure of participants' overall accuracy in interpreting the speaker's utterance. In particular, because a central hypothesis focused on the integration of mental state, action and outcome information in intentional action understanding, in this analysis we considered the study's central measures[11] together in a single analysis. Specifically, we examined the degree to which each of the study's central DVs successfully predicted scores on explanation quality. Since the Information Condition X Group interactions of central interest were largely absent across the study's primary measures for purposes of regression, we aggregated variables for each participant across information condition and participant group, yielding overall mean scores for the forced-choice belief question, the Intention component score, the Outcome component score, the mean difference between Reason and Causal history explanations, the mean difference

[11] Results from the empathy task are not included in the current analysis.

between Belief and Desire reasons, and explanatory quality. Explanatory quality was then regressed on these five variables as well as on age, IQ scores, and group membership. In this model containing eight predictors, R^2 was 0.70, with an adjusted R^2 of 0.66, indicating a large effect size (Cohen, 1988). Four of the five main study variables – Intention, Outcome, Reasons - CHRs, and Beliefs - Desires – made significant contributions to the prediction of explanatory quality (*Fs* of R change > 3.90; *ps* ≤ .05), while the forced-choice Belief question, Age, and IQ scores did not make significant contributions to prediction (*Fs* of R change < 1.61; *ps* > .21). Elimination of these latter four variables from the model yielded a final model with only four predictors, R^2 = 0.69; adjusted R square = 0.66. Bivariate zero-order correlations between all variables are reported in Table 10, and unstandardized regression coefficients (B*s*), corresponding tests of significance, and semi-partial correlations for both regression models are reported in Table 4.11.

In summary, four of the main study variables significantly predicted participants' abilities to produce a correct explanation for the utterance. Interpreting the outcome of the story as awkward but the intention as positive (nice or neutral), as well as giving mental state reason versus non-reason and belief versus desire reason explanations in response to the explanation question, all positively predicted the quality of explanation content given in response to the explanation question.

Table 4.10

Zero-order correlations for primary study variables, aggregated over condition.

	Quality	Belief	Outcome	Intention	Rea -Chr	B - D	Group	IQ
Quality								
Belief	0.20							
Outcome	0.30*	.31*						
Intention	0.32*	0.05	0.17					
Rea -Chr	0.25	0.01	-0.24	0.07				
B – D	0.80**	0.15	0.23	0.14	0.02			
Group	0.04	-0.02	0.00	0.10	0.14	0.06		
IQ	0.23	0.37*	0.12	0.13	-0.25	0.24	-0.14	
Age	-0.16	-0.12	0.07	0.19	0.15	-0.22	0.36**	-0.08

Note. Quality = Explanation Quality rating; Belief = Forced choice belief question; Outcome = Outcome component; Intention = Intention component; Rea-Chr = Difference score between Reason explanations and Causal History explanations (F.Ex coding); B – D = Difference score between Belief explanations and desire explanations (F.Ex coding); Group = Participant group (Autism vs. Control); IQ = WAIS score; Age = Chronological age.
*$p < .05$, **$p < .01$, ***$p < .001$.

Table 4.11

Results of linear regression analyses with Explanation Quality as the outcome variable, with unstandardized regression coefficients (Beta values). Significant predictors appear in bold.

	B	SE*(B)*	*T*	*P*	Semi-partial correlation
Model 1					
Rea – Chr	**0.28**	**0.08**	**3.65**	**0.00**	**0.28**
B – D	**0.61**	**0.08**	**7.47**	**0.00**	**0.58**
Outcome	**0.21**	**0.09**	**2.28**	**0.03**	**0.18**
Intention	**0.27**	**0.12**	**2.32**	**0.02**	**0.18**
Group	-0.02	0.09	-0.23	0.82	-0.02
Belief	-0.01	0.18	-0.04	0.97	0.00
IQ	0.00	0.00	0.88	0.39	0.07
Age	-0.01	0.01	-1.12	0.27	-0.09
Model 2					
Rea – Chr	**0.25**	**0.07**	**3.46**	**0.00**	**0.26**
B – D	**0.65**	**0.08**	**8.62**	**0.00**	**0.66**
Outcome	**0.19**	**0.08**	**2.29**	**0.03**	**0.17**
Intention	**0.25**	**0.11**	**2.24**	**0.03**	**0.17**

4.5 Study 7

Study 6 had been conducted with a population of French adults with ASD recruited through a local psychiatric hospital in Paris, France, a subset of whom had completed similar faux pas tasks as well as other social cognition tasks before. In Study 7, we sought to replicate

the findings of Study 6 in a population of English-speaking American adults with high-functioning autism, many of whom had only recently joined a statewide research database, and would be likely to have less familiarity with such tasks. For comparison, in Study 7 we also recruited a group of typically developing matched control participants.

4.5.1 Methods

Stimuli and measures in Study 7 were identical to those of Study 6 except for the following modifications.

Stories. Study 7 stories appeared in their original English. In addition, one of the foil stories in Study 6 (see "Charity control" in the Appendix) contained an ambiguity that was corrected for Study 7.

Measures. In order to avoid a "No" bias in responding, the forced-choice belief question in an additional two faux pas stories and one control story were reverse coded, such that the correct answer was "Yes" in 7 out of the 14 total stories presented to participants, and the correct answer was "No" in the other 7 stories.

Clinical characteristics of the sample. 22 adult participants with Autism Spectrum Disorder (DSM-V, American Psychiatric Association, 2013) were recruited in collaboration with the Rhode Island Consortium for Autism Research and Treatment (RICART), a local patient registry administered by researchers at Bradley Hospital and Brown University in Providence, Rhode Island. As a part of their participation in RICART, diagnosis for participants was

confirmed with the ADOS- 2 (Lord et al., 2000). As a part of the study, each participant first completed a verbal consent procedure administered over the telephone by a research assistant. Participants then completed the two-part online survey remotely in two sessions. The first section, lasting 1 hour and 15 minutes, included the faux pas stories and a short set of debriefing questions. The second section, lasting 20 minutes, was an online administration of a 9-item abbreviated short form of Ravens Progressive Matrices, Standard (Raven & Court, 1988, short-form derived and scored from formula provided in Bilker et al., 2012). 4 ASD participants did not complete Part II (total *N* completed = 18). 35 typically developing control participants from around the United States were recruited through flyers, word of mouth, and online advertising. These participants first completed the Ravens Progressive Matrices online task, and, then the 45-minute Faux Pas task in a separate online session. A subset of 22 of these participants were selected for inclusion in the present study such that chronological age, gender, and intelligence were matched as closely as possible with the clinical group. Demographic and clinical information is reported in Table 4.12.

Table 4.12.

Demographic information (means listed for Age and intelligence, standard deviations in parenthesis)

	Gender	Age	Intelligence
ASD	16 Men	30.59	46.06
		(14.75)	(14.35)
Control	14 Men	29.64	49.91
		(11.95)	(7.62)

Note. IQ = Scores on 9-item form of Raven's Progressive Matrices, calculated from formula provided in Bilker et al. (2012). Age = Chronological age.

4.5.2 Results and Discussion

Control Questions. As in Study 6, scores were aggregated by condition and averaged, such that each participant's score reflected the average number correct per story. As this analysis is still in progress, means for this and remaining analyses in-progress for Study 7 are marked with "XX" placeholders and standard deviations with "YY" placeholders. The mean difference in the mean number of control questions answered correctly between ASD and control participants was XX [M_{Control}= XX, SD = YY; M_{ASD} = XX, SD = YY, t(XX) = XX].

Interpretation of the utterance. For each utterance, participants had the option to check "nice," "mean," "awkward," and/or "neutral." As in Study 6, responses for each of the four variables were first aggregated across each of the 8 cells of the design, and the similarity of variance-covariance matrices was confirmed by Box's M test for homogeneity of variance-covariance matrices, M = 248.59, F(136, 5447) = 1.07, p = .28. As in Study 6, variables were aggregated across the four conditions and subsequently entered into a principle components analysis. Two orthogonal components were extracted, similar in structure to those in Study 6. The "awkward" and "nice" variables loaded most highly on the first component, while the "mean" and "neutral" variables loaded most highly on the second component. Component loadings for each of the four variables are provided in Table 4.13.

Table 4.13

Standardized component loadings for the orthogonal 2-component solution for "description of the utterance"

	Component 1	Component 2
Awkward	.93	.06
Nice	-.65	.63
Mean	-.22	-.70
Neutral	-.34	.72

As in Study 6, the two components capture the intention underlying the action (nice or neutral vs. mean, Component 2) and the outcome of the action (nice or neutral vs. awkward, Component 1). Component scores for each participant were computed in the same manner as for Study 6.

As in Study 6, the two components were uncorrelated (r = .00, *ns)*. Means and standard deviations for both components are reported in Table 4.14.

Table 4.14

Means (standard deviations) of component scores on outcome and intention

	Autism				Control			
	No Info	Belief	Desire	B & D	No Info	Belief	Desire	B & D
Outcome	0.70	0.89	0.79	1.10	0.96	0.78	0.89	1.20
	(0.63)	(0.61)	(0.70)	(0.52)	(0.53)	(0.56)	(0.62)	(0.40)
Intention	0.54	0.74	0.67	0.81	0.52	0.37	0.44	0.73
	(0.56)	(0.67)	(0.78)	(0.58)	(0.64)	(0.69)	(0.68)	(0.43)

For the outcome component, all participants endorsed "awkward" with greater frequency than other outcome descriptors when presented with both mental states vs. a single mental state [$F(1, 42) = 16.4$, $p = 00$]. This same effect held for the intention component ($F = 4.67$, $p = .04$). There were no other significant main effect contrasts for either component ($Fs(1, 42) < 1.9$, $ps > 18$), nor were there any significant Group X Condition interactions for either component (all $Fs < 2.2$, $ps > .14$). Scores in the No Information condition trended in the same direction as in Study 6, but there was no significant difference between ASD and control participants in the No Information condition, $t(42) = -1.5$, $p = .14$. A post-hoc test did indicate that ASD participants endorsed the "nice" or "neutral" descriptors with marginally greater frequency than did controls in the presence of an explicit belief, $t(42) = -1.84$, $p = .07$.

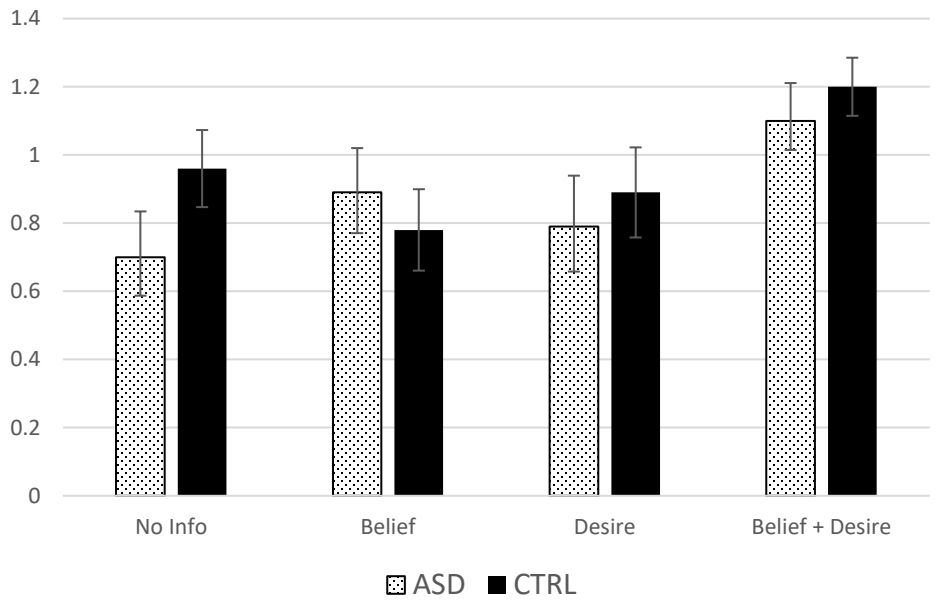

Figure 4.6. Scores on the Outcome component (where scores closer to +2 indicate endorsement of "awkward" and those closer to -2 indicate "nice" or "neutral.)"

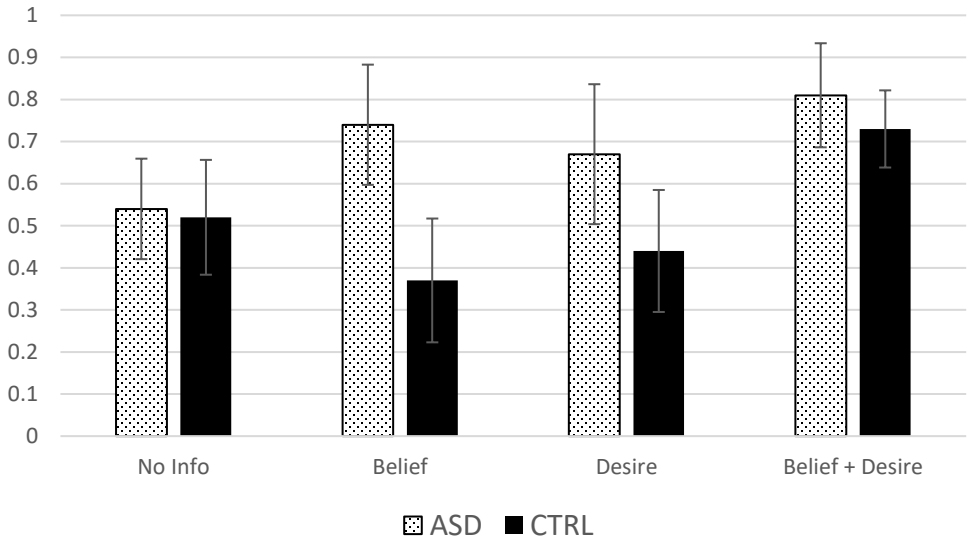

Figure 4.7. Scores on the "Intention" component (where scores closer to 2 indicate "nice" or "neutral" and scores closer to -2 indicate "mean.")

Teaching Section. Component scores were calculated in an identical fashion to Study 6 (scores are reported in Table 4.15). Planned contrasts explored differences in scores on the two components when participants were presented with full mental state information in the Teaching Section versus incomplete (no information, belief only, or desire only) information in earlier stories. For the outcome component, a planned contrast revealed no significant difference between adding both mental states to an original "No Information" story, and adding an additional mental state to a story that already contained a mental state, $F(1, 19) = 0.34$, $p = .56$. However, when a story content was presented in the Belief + Desire condition in section II after that same story content was presented in the Desire-only condition in Section I, ASD participants rated the outcome of the full-information story as *less* awkward (*M* of the Outcome component = 0.31) than the original desire-containing story (*M* of the Outcome component = 0.48), and this differed from the slight *increase* in endorsement of awkwardness (from *M* = 0.17 to *M* = 0.33) seen in response to Full Information stories that had previously been presented in the Belief-only

condition, $F(1, 18) = 10.67$, $p < .005$. There were no significant between-conditions comparisons for the Intention component ($Fs < 2.38$, $ps > .14$).

Table 4.15. *Means (standard deviations) of component scores for outcome and intention in the "Teaching" section (presented in the full information, or "Belief + Desire" conditions) and of the story contents in which those stories appeared in the belief, desire, and No Information conditions in Section I.*

	No Info	Belief + Desire (NI)	Desire	Belief + Desire (D)	Belief	Belief + Desire (B)
Outcome	0.33	0.24	0.48	0.31	0.17	0.33
	(0.42)	(0.41)	(0.35)	(0.44)	(0.31)	(0.36)
Intention	0.16	0.33	0.47	0.48	0.18	0.37
	(0.51)	(0.43)	(0.39)	(0.30)	(0.41)	(0.37)

Note: Because component scores for the teaching section are based on conditions containing a single story rather than conditions containing two stories, component scores range from -1 to 1 rather than -2 to 2.

Interpretation of the Utterance: Discussion. As in Study 6, ASD participants demonstrated a correct understanding of the speaker's intention in the No Information condition, endorsing the "nice" or "neutral" descriptors with the same frequency as did control participants. Unlike in Study 6, ASD participants' understanding of the outcome, as indicated by their frequency of endorsement of the "awkward" descriptor, was not significantly less than that of control participants (though it did trend in that direction). And also in contrast to Study 6, control participants showed no "curse of knowledge." Overall, performance of the ASD participants in the No Information condition was comparable to that of controls, providing stronger evidence than in Study 6 for the knowledge-based inferential deficit account.

In Study 7, the utterance interpretation question confirmed the successful mind-action and mind-outcome integration capacities demonstrated in Study 6. In the presence of explicit mental state information, ASD participants performed at least as well as controls on both the intention and outcome components, and even endorsed the correct descriptive terms for the intention component with greater frequency than controls in the Explicit Belief condition. Because understanding the "intention" in this question is roughly equivalent to understanding the correct valence of the speaker's desire, improved performance in the Belief condition provides specific evidence that ASD participants' understanding of the positivity of the speaker's overall intention was enhanced by an understanding that the speaker had a false belief (mind-action integration). And as in Study 6, when a story that participants had previously seen only in the Explicit Belief condition was presented with both the belief and a desire in the teaching section, ASD participants increased their endorsement of the correct descriptors for outcome, suggesting that providing an explicit desire gave additional information about the puzzle created by the conflict between the positive or neutral desire and the action's outcome (successful **mind-outcome** integration). In contrast, as in Study 6, the addition of belief information actually depressed endorsement of the correct descriptors for the outcome component, supporting the suggestion from Study 6 that a focus on the agent's belief state *prior* to performing the action may have led to diminished salience of the agent's *resulting* mental state (and recognition of awkwardness) after performing the action.

Forced-Choice Belief Question. As in Study 6, correct responses to the belief question were aggregated across story. Mean numbers of correct responses are reported in Table 4.16.

Table 4.16

Mean (standard deviation) of number correct (out of 2) on the belief question

	No Information	Belief	Desire	Belief & Desire
ASD	1.68	1.86	1.90	2.00
	(0.48)	(0.35)	(0.29)	(0.00)
Control	1.77	1.72	1.95	1.91
	(0.43)	(0.55)	(0.21)	(0.29)

Analyses conducted were the same as those in Study 6. Compared with the No Information condition, conditions in which explicit mental state information was provided (the Belief, Desire, and Belief + Desire conditions) led to a significant overall improvement in performance on the belief question [F (1, 42) = 4.70, p = .04]. There was also a marginally significant main effect of receiving both mental states in comparison to either belief or desire alone, F(1, 42) = 2.99, p = .09) and of receiving belief versus desire information, [F(1, 44) = 2.68, p = .11)]. Interactions of each of these within-subjects contrasts with the autism factor was then examined. No interactions reached significance (all Fs < 1.20, ps > .28), and as in Study 6, ASD (M = 1.68, SD = 0.48) and control (M = 1.77, SD = 0.43) participants performed similarly in the no information condition [Welch's t(41.54) = 0.66, p = 0.51].

Part II: "Teaching" section. As in Study 6, ASD subjects' improvement from the Belief, Desire, and No information conditions to the Full Information condition was examined. Three participants who did not complete the teaching section were excluded from the analysis. In both Teaching Section (Belief + Desire) stories for which participants had previously seen the same story content in the Belief only condition and for those they had previously seen in the

Desire only condition, there was no improvement (M = 0) and no variance (SD = 0). Only in Teaching Section stories which participants had previously seen in the No Information condition was there any improvement from the original stories to the Belief + Desire stories (M = 0.21, SD = 0.42). This improvement differed significantly from 0, $t(18) = 2.19$, $p = .04$.

Forced choice belief question: Discussion. As in Study 6, ASD and control participants performed comparably in the No Information condition, providing further evidence for the knowledge-based inferential deficit account. In contrast to Study 6, both groups showed consistent improvement in response to explicit mental state information, but as in Study 6, both groups' performance improved relative to the No Information condition in the presence of explicit desire *and* belief information, suggesting not only that ASD participants' understood the forced-choice belief question, but that they were able to use the desire information provided in the story to successfully achieve *mind-outcome* integration, detecting the intentional action puzzle and resolving it with the appropriate belief content.

Responses to the Explanation question: Conceptual Form.

F.Ex Coding Reliability. In Study 7 there were a total of 14 stories each for 22 control participants, 17 stories each for 20 ASD participants, and 14 stories for 2 ASD participants. For reliability purposes, two coders classified 72 open-ended responses from Study 7. Coders reached 95% agreement on distinguishing reason explanations from causal history explanations ($\kappa = .72$), 100% agreement on distinguishing belief reasons from desire reasons ($\kappa = 1.0$), and 100% agreement on the application or non-application of mental state markers (e.g., "he thought, she wanted") to reason explanations, ($\kappa = .73$). Agreement on classification of trait vs. nontrait

causal history explanations was 100% (κ = .78). On a larger group of 113 items including the original 72, coders reached 100% agreement on the codability of explanations (κ = .89). All remaining items were classified according to the F.Ex coding scheme (Malle, 1998/2010) by a single coder blind to participant group and hypotheses.

Results. As before, means reflect the number of explanations given *per story,* aggregated across story and within information condition. Means for Reason vs. CHR explanations are displayed in Figure 9, and means and standard deviations for Reasons vs. CHRs, as well as Belief vs. Desire reasons and the use of mental state markers are displayed in Table 16. As in Study 6, mixed (Information X Autism) ANOVAs were performed on difference scores for Reasons - CHRs, Beliefs - Desires, Unmarked beliefs - Marked beliefs, and Non-trait CHRs – Trait CHRs.

Reasons and Causal Histories. Control and ASD participants gave comparable numbers of reason and causal history explanations for stories that did and did not contain explicit mental state information [$F(1, 42)_{\text{interaction}} = .2.40$, $p = .13$], as well as comparable numbers of these explanation types in the presence of a single versus both mental states, [$F(1, 42)_{\text{interaction}} = 1.15$, $p = 0.29$). In contrast to Study 6, there was no difference between ASD and control participants in the relative use of reasons and causal histories in response to explicit belief versus explicit desire information [$F(1, 42)_{\text{interaction}} = .087$, $p = 0.77$).

Belief and Desire Reasons. ASD and control participants gave comparable numbers of beliefs and desires across the four mental state information conditions, $Fs_{interaction} < 0.97$; $ps > .33$.

Beliefs: Mental State Markers. Compared with control participants, ASD participants provided a greater number of mental state markers for beliefs in the No Information condition relative to the explicit mental state information conditions, $F(1, 22)_{interaction} = 4.74$, $p = .04$, although both groups gave fewer markers overall in the No Information condition than in the explicit mental state information conditions, $F(1, 22) = 0.05$, $p = .05$. This finding contrasted with the finding in Study 6, in which this trend was reversed.

Causal histories: Traits. There were as few as four valid participants per cell for purposes of ANOVA for difference scores; therefore, formal analyses are not reported. As in Study 6, the raw number of traits for all participants was small in all conditions for both groups (e.g., for No Information, $M_{Autism} = 0.10$, $SD = 0.26$; $M_{Control} = 0.15$, $SD = 0.34$; for Explicit Belief, $M_{Autism} = 0.13$, $SD = 0.29$; $M_{Control} = 0.28$, $SD = 0.58$).

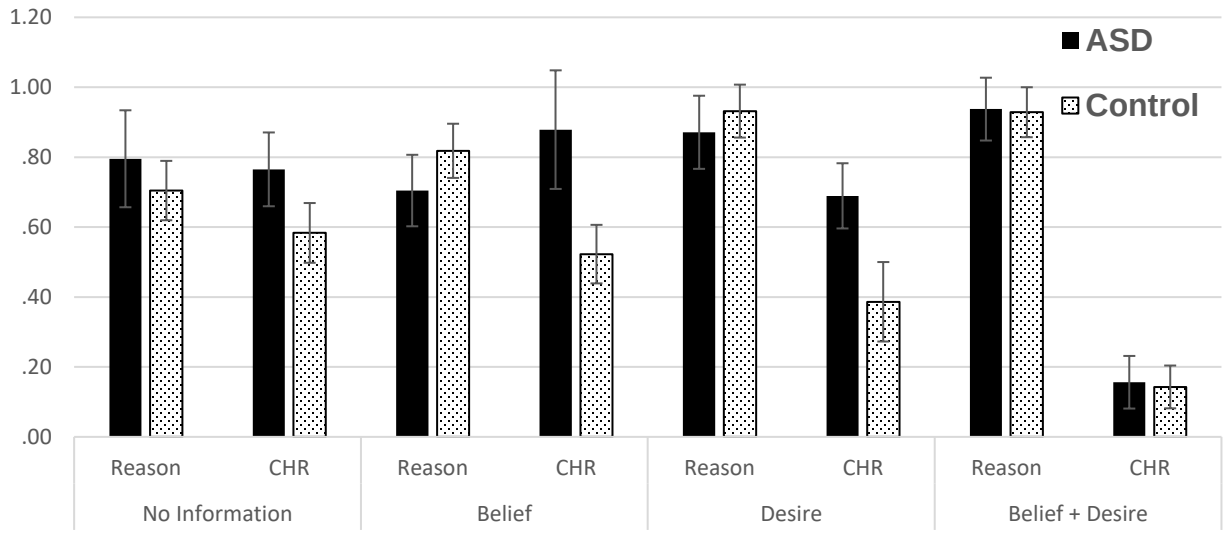

Figure 4.8. Number of Reason and CHR explanations (per behavior) given in each of the four information conditions.

Table 4.17

Mean (standard deviation) number of explanations per behavior in each of the four information conditions. Top panel: Comparisons of Reasons vs. CHRs. Middle panel: Comparison of Belief versus desire reasons. Bottom panel: Comparison of marked versus unmarked beliefs.

	No Info		Belief		Desire		Belief + Desire	
	Reason	CHR	Reason	CHR	Reason	CHR	Reason	CHR
ASD	0.80 (0.65)	0.77 (0.50)	0.70 (0.48)	0.88 (0.80)	0.87 (0.49)	0.69 (0.44)	0.70 (0.50)	0.70 (0.50)
Control	0.70 (0.40)	0.58 (0.40)	0.81 (0.36)	0.52 (0.39)	0.93 (0.36)	0.39 (0.53)	0.93 (0.28)	0.27 (0.43)
	Belief	Desire	Belief	Desire	Belief	Desire	Belief	Desire
ASD	0.83 (0.49)	0.40 (0.60)	0.97 (0.21)	0.17 (0.42)	0.65 (0.52)	0.48 (0.48)	0.94 (0.36)	0.16 (0.30)
Control	0.78 (0.35)	0.28 (0.39)	0.95 (0.21)	0.10 (0.26)	0.72 (0.43)	0.20 (0.37)	0.92 (0.33)	0.14 (0.28)
	Marked	Unmarked	Marked	Unmarked	Marked	Unmarked	Marked	Unmarked
ASD	0.91 (0.51)	0.18 (0.39)	0.83 (0.42)	0.19 (0.39)	1.00 (0.42)	0.10 (0.28)	1.06 (0.25)	0.00 (0.00)
Control	0.68 (0.41)	0.32 (0.41)	0.95 (0.15)	0.07 (0.18)	1.00 (0.34)	0.08 (0.26)	1.05 (0.22)	0.00 (0.00)

Teaching section. Mean number of reasons, causal histories, beliefs, desires, and each type of marked reason explanation were calculated for Belief + Desire stories that followed the same story contents in each of the incomplete information conditions. Because each participant responded to only one Belief + Desire story for each of the incomplete information conditions and participants tended to give only a single explanation type, the *N* of valid difference scores was too small (*N* = 9*)* for any explanation parameter to perform formal statistical analyses.

Quality Coding. Quality coding of responses to the explanation question was performed on all Study 7 Faux Pas stories according to the guidelines established for Study 6.

Reliability. 415 responses were classified by two coders. Overall agreement on all explanation types was X% (κ = X). Consensus on X items (X% of all items) could not be reached and these items were excluded from analysis.

Quality coding analysis.

"Don't know" responses. The number of "Don't know" responses was negligible, with ASD participants giving a total of 6 (1.4%) such responses and Control participants gave a total of 1 (.02%). For the remaining analyses, don't know responses were treated as missing values.

For explanation quality ratings, three planned contrasts examined main effects for the information factor. Means for each condition are reported in Table 4.18. In contrast to Study 6, compared with the No Information condition, there was an overall improvement in conditions in which explicit mental state information was provided, $F(1, 42) = 4.48$, $p = .04$. And like in Study 6, in the presence of both mental states, participants provided higher quality responses than they did in the presence of a single mental state, $F(1, 42) = 3.54$, $p = .07$. There was no difference in overall response quality between the belief and desire conditions, $F(1, 42) = 1.97$, $p = .17$). As in Study 6, there were no significant interactions between autism and information condition (all $Fs < 1.26$, $ps > .27$). There was also no difference between control ($M = 2.09$, $SD = 0.72$) and ASD ($M = 1.90$, $SD = .87$) participants in the No Information condition, $t(42) = 0.76$, $p = .45$.

Table 4.18

Mean (standard deviation) of quality score on the explanation question

	No Information	Belief	Desire	Belief & Desire
Autism	1.91	2.39	2.09	2.36
	(0.87)	(0.67)	(0.91)	(0.69)
Control	2.09	2.14	2.07	2.41
	(0.72)	(0.68)	(0.74)	(0.65)

Negative Intentions. Two coders classified every response as either citing or not citing an incorrect negative intention, and agreed XX% on distinguishing an incorrect negative intention response from a neutral or correct response (κ = XX).

Lack of knowledge. A single coder familiar with the structure of the responses performed these classifications. Mean counts for correct (perfect score) belief contents and correct lack of knowledge contents will be reported in Table XX.

Explanations of behavior: Discussion. As in Study 6, ASD and control participants provided similar numbers of reason and causal history explanations in the No Information condition. In addition to making use of similar concepts in their explanations, ASD and control participants provided explanations of similar quality in the No Information condition, providing further evidence for the knowledge-based inference account. In spite of these continuities with Study 6, one clear discontinuity emerged. Curiously, the use of mental state markers in the No

Information condition trended in the opposite direction from Study 6, with ASD participants using more marked beliefs relative to unmarked beliefs compared with controls.[12]

Study 7 provided even stronger evidence than Study 6 for ASD participants' intact ability to integrate mental state information for intentional action understanding. In addition to showing clear improvement, comparable to that of controls, in the presence of both explicit mental states, the scores of both groups also improved from the No Information condition to the three explicit information conditions. ASD participants also sustained this improvement in the desire condition, showing further evidence for mind-outcome integration by generating the correct belief.

Empathy Question. Post-testing for the empathy question for Study 7 is underway. Means for each of the four conditions (M = XX, SD = YY) will be reported in Table XX.

Summary analysis: Regression of main variables on explanatory quality. As in Study 6, the Information Condition X Group interactions of central interest were largely absent across the study's primary measures. Thus, for purposes of regression, we once again aggregated variables for each participant across information condition, yielding overall mean scores for each participant for the forced-choice belief question, the Intention component, the Outcome component, the mean difference between Reason and Causal history explanations, the mean difference between Belief and Desire reasons, and explanatory quality.

As in Study 6, explanatory quality was regressed on the five central variables of interest (forced-choice belief, Intention, Outcome, Reason-Causal history difference score, and Belief-

[12] Although no obvious explanation for this difference is apparent, one could speculate it is due to cross-linguistic differences in the use of mental state terms (between French and English), or to some interaction between language delays in ASD participants with cross-linguistic differences.

Desire difference score) as well as on age, intelligence scores, and group membership. In this model containing eight predictors, R^2 was 0.52, with an adjusted R^2 of 0.37. The Belief – Desire difference score made a significant contribution to the prediction of explanatory quality (*F* of R change = 22.48; *p* = .00), and the Reason – Causal History difference score made a marginally significant contribution to the model (*F* of R change = 2.98, *p* = .10). None of the remaining predictors, including Age and intelligence, made significant contributions to prediction (*Fs* of R change < 1.87; *ps* > .18). Elimination of these six variables from the model yielded a final model with only two predictors, R^2 = 0.47; adjusted R square = 0.43. Bivariate zero-order correlations between all variables are reported in Table 19, and unstandardized regression coefficients (B*s*), corresponding tests of significance, and partial correlations for both regression models are reported in Table 4.20.

In summary, two of the main study variables predicted participants' abilities to produce a correct explanation for the utterance. Giving mental state reason versus non-reason and belief versus desire reason explanations in response to the explanation question both positively predicted the quality of explanation content given in response to the explanation question.

Table 4.19

Zero-order correlations for primary study variables, aggregated over condition.

	Quality	Belief	Outcome	Intention	Rea -Chr	B - D	Group	Intell.
Quality								
Belief	0.25							
Outcome	-0.30*	.00						
Intention	-0.15	0.01	0.10					
Rea -Chr	0.06	0.12	0.26	0.06				
B – D	.61**	0.16	-0.50**	-0.32*	-0.29			
Group	0.04	0.06	0.08	0.38*	-0.24	-0.14		
Intell.	0.34	0.42*	-0.25	-0.26	0.24	0.26	-0.14	
Age	-0.42**	-0.25	0.07	0.11	-0.03	-0.39*	.364**	.165

Note. Quality = Explanation Quality rating; Belief = Forced choice belief question; Outcome = Outcome component; Intention = Intention component; Rea-Chr = Difference score between Reason explanations and Causal History explanations (F.Ex coding); B – D = Difference score between Belief explanations and desire explanations (F.Ex coding); Group = Participant group (Autism vs. Control); Intell. = Ravens Short Form score; Age = Chronological age.
*$p \leq .05$, **$p < .01$, ***$p < .001$.

Table 4.20

Results of linear regression analyses with Explanation Quality as the outcome variable, with unstandardized regression coefficients (Beta values). Significant predictors appear in bold.

	B	SE*(B)*	*T*	*P*	Semi-partial correlation
Model 1					
Rea – Chr	0.267	0.164	1.628	0.116	0.225
B – D	0.624	0.210	2.973	0.006	0.411
Outcome	-0.050	0.199	-0.254	0.802	-0.035
Intention	-0.141	0.213	-0.663	0.513	-0.092
Group	0.196	0.158	1.240	0.227	0.171
Belief	0.132	0.517	0.255	0.800	0.035
Intell.	0.002	0.007	0.339	0.737	0.047
Age	-0.004	0.007	-0.604	0.551	-0.084
Model 2					
Rea – Chr	0.229	0.133	1.725	0.095	0.227
B – D	0.709	0.138	5.146	0.000	0.677

4.6 General Discussion

In this study, participants were presented with "faux pas" stories that depicted a character who, due to a false belief, unintentionally insults another character. Previous studies have shown that adolescents and adults with autism struggle to identify the mental states underlying

such actions with unintended results, at times incorrectly assuming that the character acted on the basis of a true belief or a negative intention.

The present study explored two possible explanations for these struggles. On one account, the struggle is inferential. This inferential deficit may be explained by one of two hypotheses. First, it could be that high-functioning individuals on the autism spectrum are unable to infer beliefs and desires from contextualized and naturalistic situations – that is, they have a theory of mind deficit that persists from childhood into adulthood. Alternately, ASD adults' inability to draw the appropriate inferences from the stories in the past may be due to struggles with the flexible use of general knowledge. This second hypothesis has plausibility because the background knowledge needed to draw appropriate belief and desire inferences may be unavailable in faux pas stories found in previous studies.

Beyond their inferential struggles, ASD adults may struggle on the faux pas stories in Zalla et al. (2009), and on other similar "advanced" theory of mind tasks (e.g., Happé , 1994), not primarily or exclusively because of a deficit in the ability to draw belief and intention inferences, but for a different reason: they may be incapable of *integrating* these disparate mental state and reality representations together into a holistic understanding of the causes, effects, and meaning of intentional action.

4.6.1 Inference

To test the inference account, we modified previous faux pas stories by presenting participants with ample background information from which to draw a mental state inference. When presented in the No Information condition, these stories were structurally similar to previous faux pas stories but were more explicit in offering background information. The

comparison between typically developing and control participants in this condition could serve as a test of whether documented deficits are due only to continuing impairment in belief and desire inference, or whether they were due to previous stories' insufficient background affordances for mental state inferences. The specific background we presented was intended not only to suggest that there was something that the speaker did not know, but also to provide an affordance for a plausible inference of the actual content of the speaker's belief. For example, one story describes Sandeep, who has lived in New Jersey his entire life, a state in which it is illegal to pump one's own gas. Sandeep moves to Pennsylvania, and when he sees an older man near one of the pumps (really just another customer), he asks him to pump his gas for him. Because the social perceiver knows that Sandeep is used to having his gas pumped for him by attendants, the story affords the inference that Sandeep falsely believes that the man is the attendant.

Across almost all study measures in two culturally distinct populations, high-functioning adults on the autism spectrum performed similarly to typically developing control participants in response to stories like these in the No Information condition. They were equally able to correctly identify the speaker's knowledge state in a forced-choice question, and to identify the speaker's intention as benign. Overall, even in the absence of explicit mental state information, ASD participants showed a remarkable ability to provide the correct answer among provided choices and to provide the correct mental state or other background content in a similar conceptual form to that used by control participants. They were capable of this even in response to an open-ended question for which any mental state inferences they provided reflected their own self-generated representation of the behavior's outcomes. The only sustained difference between ASD and control participants in the No Information condition was in the two groups'

endorsement of the "awkward" descriptor versus other descriptors of the action's outcome in Study 6, but this finding was not fully replicated in Study 7.

In particular, ASD adults' consistently high performance on both forced-choice and open-ended identification of the speaker's false belief calls into question the use of the faux pas task as a targeted measure of "advanced" theory of mind abilities in high-functioning adults with autism. It suggests that ASD adults' inferential deficits on previous "advanced" theory of mind tasks, and especially on the faux pas task, may be due not to any enduring deficit in the ability to draw mental state inferences from complex behaviors, but due to specific features of those tasks that made commonly used compensation strategies ineffectual. Specifically, beyond requiring mental state inferences, the sparse background information provided in these tasks required participants to generate novel connections between social roles and the behaviors most probably associated with those behaviors, such as the idea that uncles do not insult their nieces completely unprompted. When the background information needed to make the most probable mental state inference is made explicit in the stories without providing the mental state inference itself, however, performance on false belief inference as well as understanding of the positive valence of the agents' intention improves (from previous versions of the faux pas stories) to the same level of control participants. It should be noted, however, that although about half of the participants in Study 6 had previously completed earlier versions of the faux pas stories (and exhibited the earlier-discussed struggles in faux pas understanding), that the current data do not reflect a direct, within-subjects comparison of stories enriched with inferential affordances with stories unenriched in this manner. Thus the current findings on inference should be replicated including a within-subjects comparison of enriched versus unenriched stories.

Our stories were not devoid of inferential challenges; they required mental state inferences that drew on a variety of affordances: appreciating, for example, that if a city is politically liberal, a person might (mistakenly) come to believe that his squash partner (who lives in that same city) is liberal as well, or that a parent who is short and dresses nondescriptly might be wrongly perceived by others as a child if she is at a school. Of course, our stories point participants toward this information by presenting the background premise explicitly and at an appropriate place in the narrative. Critically, however, ASD participants are able to draw connections between these pieces of information and the character's utterance so as to generate mental states, both in response to explicit questions and in response to the less constrained prompt to explain why the speaker said what she said. The idea that ASD individuals may understand certain utterances as caused *primarily* by the same (correct) mental states generated by controls suggests that, in the presence of the right inferential affordances, they are capable of detecting when a behavior's apparent meaning (e.g., a mean remark) is actually the result of a character's underlying beliefs and desires. They are able to richly represent the world by inferring a range of mental states, a task that requires generating a unique mental state content for every unique story, and requires more than "hacking out" a rule-based, non-mentalistic solution to structured false belief problems (Happé, 1995). These results also suggest support for the hypothesis that, insofar as high-functioning individuals on the autism spectrum still struggle with inferring others' beliefs and desires in naturalistic situations, this may be due more to their failure to converge with typically developing adults on correctly identifying the information most 'relevant' to the context than it does with any explicit conceptual or inferential competence (Apperly, 2011).

4.6.2 Integration

Beyond the inferential abilities demonstrated in the No Information condition, ASD and participants also demonstrated successful mind-action integration at the same rate as control participants, using mental state information to produce a correct description of the intentional action as caused by a positive intention and a false belief. They also demonstrated successful mind-outcome integration, recognizing that, distinct from the intention, the resulting outcome was negative in valence (offensive to the listener and awkward for the speaker).

Although ASD individuals already demonstrated significant integration capacities in the No Information condition, neither they nor control participants performed at ceiling on forced-choice measures of belief understanding or on the quality of their explanations. Thus the conditions in which belief alone, desire alone, or both mental states were explicitly presented still provided an additional opportunity to examine the abilities of ASD adults to use this additional information to enhance their understanding of action.

In the explicit belief condition, ASD participants were as capable as control participants at mind-action integration, and even endorsed the speaker's positive intention at a higher rate than did control participants in Study 7. They also successfully integrated the explicit belief information with the outcome (mind-outcome integration), providing high-quality open-ended explanations that cited both the content of the belief and acknowledged its falsity, and recognizing the unintended outcome of the action – offense of the listener – as accurately as did control participants. Although it was manifest in different measures on Study 6 and Study 7, when presented by itself, belief information seemed to have special salience for ASD participants' evaluation of the actor's intention, leading them to endorse the positive intention even more strongly than control participants (Study 7), and to provide an unusually high number

of reason explanations relative to controls, and compared to the explicit desire condition (Study 6).

In the explicit desire condition, ASD participants exhibited both an intact ability to recognize the desire's relevance to the speaker's intention in performing the action (successful mind-action integration) relative to controls, as well as an ability to infer the speaker's belief based on a detected inconsistency between the desired outcome and the actual outcome (successful mind-outcome integration). Evidence of the intact ability to integrate mind with outcome is particularly compelling in the explicit desire condition, in which the agent's desired outcome and the actual outcome of the action are in direct conflict. This evidence was especially strong in Study 7, in which ASD participants not only performed comparably to controls, but exhibited improvement in explanatory quality the desire condition along with controls.

Furthermore, in the presence of an explicit belief combined with an explicit desire, both ASD and control participants in both studies provided higher-quality explanations than in the presence of a single mental state. This finding suggests that ASD individuals have a capacity for mind-outcome integration that is comparable to that of controls in taking into account conceptually relevant information. Like controls, they are capable of understanding how an action that is motivated by a positive desire leads to a negative outcome due to the speaker's false belief, and the presence of a desire improves understanding over and above the presence of a belief by itself.

In addition, in the Teaching section, when a desire was *added* to a story that had previously been presented in the belief condition, in both studies ASD participants showed an improvement in mind-outcome integration by demonstrating an increased awareness of the mental state the speaker would have after the action was completed. The failure to sustain this

improvement in response to the addition of a belief to a desire-containing story suggests that for ASD participants, there may be some cost in emphasizing an agent's belief state *prior* to acting if that state changes as a result of the action. Because the Teaching section was only presented to ASD participants, the present data cannot speak to whether this differential improvement is unique to them.

In addition to their performance in each condition individually, the overall ability to provide accurate explanations was, across the two populations in both studies, successfully predicted by the conceptual form of the explanations (beliefs specifically, and mental state reasons more generally). This finding contrasts with previous work (Klin, 2000; Happé, 1994; Abell, 2000; Castelli et al., 2002) indicating that while high-functioning individuals with ASD can provide mental states as explanations for intentional actions, that these mental states are often incorrect or irrelevant. Whether or not they received explicit belief information, ASD participants cited accurate belief reasons that served to link the speaker's intention to her action, and they also recognized as well as controls that the outcome of this action was emotional distress for the listener.

4.6.3 Intention Understanding and Moral Judgment: Two Distinct Phenomena

Moral judgment of accidental harms is often described as simply a special case of theory of mind understanding: a sort of "stress test" for theory of mind competence in very high-functioning individuals who have otherwise achieved it through compensation strategies (Moran et al., 2011, p. 2691). Specifically, this is a test of ASD adults' abilities to use belief understanding to exculpate an agent who unintentionally brings about an action with a morally bad outcome. But, distinct from some previous findings, which may point to deficits in ASD

adults in the moral judgment of accidental harms, the present studies suggest that ASD adults are fully capable of integrating mental state information with information with outcome information for intentional action understanding. How can these findings be reconciled?

While moral judgment certainly makes use of intention (or lack-of-intention) information, ASD participants' moral judgments do not differ from controls' under all circumstances of inconsistency between mental states and outcomes. Specifically, ASD individuals do *not* differ from TD individuals in their moral judgment of attempted harms (Moran et al., 2011). A divergence between judgments for attempted harms and accidental harms is consistent with findings in typically developing children, who rely first on a harm-based process of moral judgment, and increase their facility in the use of intention information for the judgment of accidental harms between the ages of four and eight (Cushman et al., 2013). Thus, ASD individuals may, like typically developing individuals, make moral judgments that are driven by the initial detection of a harmful or other norm-violating event.

What, then, could account for the difference between ASD individuals and typically developing controls in the accidental harm condition in Moran et al. (2011) if not straightforward and failure on the part of ASD individuals to integrate mental state information?

As a process, moral judgment, and the exculpation of accidental outcomes, involves more than the integration of mental state information. First, ASD individuals may recognize that the outcome is accidental, and may even be able to use this intentionality information to mitigate the harshness of their moral judgments to some degree. This capability is still consistent with the Moran et al. (2011) finding, which is qualitatively the same as that of control participants (judged as more permissible than an intentional harm, and less permissible than a neutral intention/neutral outcome). The explanation for the subtle quantitative difference found in the

accidental harm condition in Moran et al. (2011) may lie elsewhere. For example, the path model of blame (Malle, Guglielmo, & Monroe, 2014) suggests that in deciding how much blame an agent deserves for an accidental, unintentional outcome, social perceivers must also consider the agent's obligation to prevent the outcome. For individuals with ASD, explicitly provided mental states may still exert considerable mitigating force in the presence of a negative outcome, but when assessing blame, they may still appeal, rather more inflexibly than TD individuals, to fixed rules about the agent's general obligation to prevent negative outcomes (e.g., "One should always go out of one's way to prevent another person's death,"). This use of norm information in one specific node in the blame process may lead to somewhat harsher blame judgments than provided by controls for accidental harms. The use of inflexible rules by ASD individuals for determining the agent's obligation to prevent is also consistent with the literature on ASD individuals' struggles to draw flexibly on their general knowledge about social roles and events (e.g., Loth et al., 2008; 2011). This appeal to fixed rules is perhaps especially true in the case of the Moran et al. (2011) findings, which do not actually elicit blame judgments, but *permissibility* judgments. Such judgments are likely to elicit an even more direct assessment of a behavior's adherence to norms (Malle et al. 2014). Viewed in this way, it is actually fairly impressive that ASD individuals still appear to mitigate their moral judgment at all based on the mental state information provided in the accidental harm condition.

4.6.4 Concepts, Inferences, and Moral Judgment

More generally, the present findings suggest a more nuanced picture of the relationship between concepts and inferences in interpreting previous findings on theory of mind, moral judgment, and their interaction in studies of high-functioning autism. Although there is abundant

evidence that ASD individuals often fall short in generating the correct contents of mental states when presented with insufficient inferential affordances, they do not spontaneously and indiscriminately rely on normative rules to understand all behavior. Rather, ASD individuals provide normative rules only in response to a prompt to explain why something is "wrong" or "bad" (Shulman et al., 2012; Zalla et al., 2009; Zalla et al. 2011).

In particular, the present findings suggest that ASD participants provided normative rules and negative intentions in their explanations for the speakers' faux pas in Zalla et al. (2009) for two reasons. First, in the absence of sufficient inferential affordances, ASD individuals were unable to generate uniformly correct mental states. Although they did provide a small number of negative intentions, they also provided correct mental states that reflected an incomplete, but not incorrect, understanding of the scenario; for example, citing the speaker's positive desire, but failing to grasp the speaker's false belief or the negative emotional impact of the outcome. Second, importantly, the explanation question in Zalla et al. (2009) did not actually elicit an explanation for the speaker's behavior, but an explication of the speaker's *norm violation* (e.g., "Why was it wrong?"). Given that ASD individuals had, as evidenced by their low performance on detecting the character's belief, an incomplete representation of the speaker's mental states to begin with, it is not surprising that they sometimes responded to this question by providing a negative intention. However, this response pattern does not reflect a lack of conceptual understanding of intentional action. Rather than relying exclusively on the diagnostic power of norm-violations to indicate an agent's intention, in the presence of sufficient inferential affordances in the present studies, individuals with high-functioning ASD demonstrated the ability to integrate mental state information for intentional action understanding: a clear conceptual competence. This divergence of findings between the domain of moral judgment and

that of intentional action understanding suggests that, as in typically developing individuals, moral judgment and intentional action understanding are distinct phenomena.

4.6.5 Future Directions: Concepts and Processes in Intention Understanding

The current findings suggest that high-functioning individuals on the autism spectrum have an intact explicit concept of intentional action. According to this concept, in order for an action to bring about a certain outcome intentionally, the agent must have *both* a desire for that outcome and belief that the action will bring about that outcome (Malle & Knobe, 1997). This model dictates that a judgment about the intentionality of a particular action cannot be reached in the absence of one of these elements.

Although this may be true conceptually, recent evidence suggests that for typically developing social perceivers, the online process of reaching intention inferences in the presence of conflicting, explicit mental state information is more complex. Consistent with the earlier developmental emergence of the desire concept (e.g., Wellman & Wooley, 1990), desires may take precedence for purposes of a perceiver's online intention inference, with the mere presence of a desire often being enough to license an initial intention inference for typically developing individuals (Malle & Holbrook, 2014; Haigh & Bonnefon, 2015). If it is also revealed that the agent *lacked* the belief that the chosen action would lead to the desired outcome, however, the perceiver then takes this information into account and adjusts his inference about the agent's chosen action, though at considerable cognitive cost (Haigh & Bonnefon, 2015). This adjustment may also be short lived, because it competes against the social perceiver's "cursed knowledge": the awareness, for example, that performing a certain action *would* fulfill the character's desire, even if the character is unaware of this.

There are several applications of these findings to future work with high-functioning autistic populations. Although ASD adults demonstrate a clear ability to integrate explicit mental state information into an accurate understanding of the concept of intentional action, the processes by which they use this information to reach their inferences about the agent's intention remain unexplored. In the version of the faux pas task presented here, an eye-tracking-while-reading paradigm similar to that used in Haigh & Bonnefon (2015) could provide precise information about the priority that ASD individuals place on different types of mental state and outcome information in reaching the inference that an outcome was brought about unintentionally. An intriguing possibility is that ASD individuals may be slower overall in combining this information to reach intention inferences that take into account both belief and desire information, but they may be less inclined than typically developing individuals to form any intention inferences at all in the absence of clearly stated belief information. Thus, while their inferential process may be less efficient, they may, at least in certain contexts, be less prone to the "curse of knowledge" than their typically developing peers.

CHAPTER 5.

Conclusion

According to the concept of intentional action, a person brings about an outcome *intentionally* if he or she has a desire for that outcome, and a belief that her chosen action will lead to that outcome. For many commonly encountered actions, the individual components of this concept contain redundant information – the person's desire accurately specifies the outcome that is actually realized by the action, and his belief accurately describes the action's status as a means to the achievement of this outcome. This informational redundancy means that the concept of intentional action may often supply very little beyond the understanding that an action that is described has fulfilled its goal. In fact, as long as an action fulfills its apparent goal, this action can often be considered independently from any specific mental states the agent may have had. Specifically, it can be considered from the "teleological stance" discussed in Chapter II, and the "cause" of the action is regarded as the (realized) goal itself. Only when a goal-directed action diverges from the fulfillment of its specific, presumed goal are the agent's specific mental states potentially important to the social perceiver's understanding of the agent's action.

In this dissertation, I have explored how social perceivers reach an understanding of intentional action in response to behavioral puzzles: cases in which an action does not fulfill its expected goal. In Chapters 2 and 3, I presented five studies that demonstrated the specific role of the concept of intentional action and its component mental states in social perceivers' understanding of puzzling actions. Even though causal background, or causal history explanations, can imply the *types* of goals particular agents might pursue (e.g., a strange person might have the goal of acting strangely), or about the types of goals agents in general might pursue in particular situations (e.g., a person in a stage play might aim to act out the strange plot

of the playwright), agents primarily achieve understanding by specifying the *particular* mental states on a specific agent's mind at a particular moment, in a particular context. Social perceivers use mental state concepts – beliefs and desires – to recombine knowledge from disparate domains in a way that sheds light on the specific, and often idiosyncratic, information that an agent may have considered before deciding to perform a particular action. Furthermore, as demonstrated in Chapter 3, as concepts, belief and desire guide the social perceiver to search for distinct types of information that shed light on this choice to act.

In Chapter 4, I presented two studies examining the ability of individuals on the autism spectrum, like typically developing individuals, to use the concept of intentional action to reach an understanding of behavioral puzzles. In contrast to prior work pointing to deficits in understanding action in terms of mental states, individuals on the autism spectrum demonstrated a clear ability to integrate explicitly presented information using the concept of intentional action.

5.1 The Role of Prior Knowledge in Mental State Inferences: Typical Development and Autism

While both typically developing and autistic individuals appear to have an intact concept of intentional action that they can apply to the understanding of behavioral puzzles, the two lines of work suggest that typically developing and autistic individuals may reach mental state inferences via distinct pathways. While typically developing individuals appear to be able to use mental state concepts to draw productively on their existing knowledge structures to generate the specific mental states a particular agent has on his mind at a particular time, individuals on the autism spectrum demonstrate somewhat more limited abilities. Notably, in the presence of

sufficient inferential affordances, individuals with ASD are capable of inferring an agent's false belief based on an intentional action that leads to an unintended outcome. They can do this even when there is no simple rule or compensation strategy (e.g., "perception leads to knowledge") to appeal to, and they must draw novel connections between their knowledge about the features of a situation and the effect it could have on another person's belief state. However, the present studies, which presented ASD participants with specific affordances to facilitate the inference of the character's mental states, stand in contrast to previous work (Zalla et al., 2009), in which autistic participants performed poorly in identifying the mental states underlying a characters' action in the absence of such affordances. This poor performance may have resulted not from a deficit in mental state understanding per se, but rather, from ASD individuals' struggles to recognize the broader inferential relevance of otherwise unrelated knowledge structures to the most probable interpretation of action (e.g., recognizing that Uncles typically do not insult their nieces without provocation).

Taken together, the work presented here on typically developing and autistic adults suggests that, even while the productive use of general knowledge may be separable from mental state inference abilities, the use of general social knowledge plays an important role in the correct interpretation of others' intentional actions. While typically developing individuals are capable of drawing broad inferences from their general knowledge that point them toward the correct mental state inferences, individuals on the autism spectrum may need additional inferential cues – explicitly stated pieces of general knowledge -- to guide them toward the correct mental state inferences. Yet, as Chapters 2 and 3 demonstrate, in the face of the most challenging behavioral puzzles, the most highly flexible use of mental state concepts in typically developing individuals– in processes such as mental simulation – may draw precisely on the

ability to generate novel combinations of general knowledge. Further study of the mechanisms of explicit mental state inferences in high-functioning adults on the autism spectrum may shed light on the boundaries between mental state inference and the productive use of general knowledge.

BIBLIOGRAPHY

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1–16. http://doi.org/10.1016/S0885-2014(00)00014-9

Aldridge, M. A., Stone, K. R., Sweeney, M. H., & Bower, T. G. R. (2000). Preverbal children with autism understand the intentions of others. *Developmental Science*, *3*(3), 294–301. http://doi.org/10.1111/1467-7687.00123

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders DSM-IV-TR fourth edition (text revision). Washington, DC.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC.

Anscombe, G.M (1965). *Intention.* Oxford: Blackwell.

Apperly, I.A. (2011*). Mindreaders*: *The Cognitive Basis of "Theory of Mind".* Hove, UK: Psychology Press.

Astington, J. W, &. Gopnik, A. (1991). Developing understanding of desire and intention. In A. Whiten (Ed.), Natural theories of mind: Evolution, development and simulation of everyday mind-reading (pp. 39-50). Oxford, England: Blackwells. Astington, J. W., Harris, P. L., & Olson, D.

Baird, J. & Astington, J. W. (2005). The development of the intention concept: From the observable world to the unobservable mind. In R. R. Hassin, J. S. Uleman & J. A. Bargh (Eds.) *The new unconscious* (pp. 256-276). New York: Oxford University Press.

Baird, J. A., & Moses, L. J. (2001). Do Preschoolers Appreciate That Identical Actions May Be Motivated by Different Intentions? *Journal of Cognition and Development*, *2*(4), 413–448.

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

Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2014). Joint inferences of belief and desire from facial expressions. In *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society* (pp. 923-928).

Banerjee, R., & Watling, D. (2005). Children's understanding of faux pas. *Hellenic Journal of Psychology, 2*(1), 27-45.

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. New York: Oxford University Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46. http://doi.org/10.1016/0010-0277(85)90022-8

Baron-Cohen, S., O'Riordan, M., & Jones, R. (1999). A new test of social sensitivity: Detection of faux pas in normal children and children with Asperger syndrome. *Journal of Autism and Developmental Disorders 29,* 407-418.

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders, 29*(5), 407–418. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10587887

Begeer, S., Malle, B. F., Nieuwland, M. S., & Keysar, B. (01/2010). Using Theory of Mind to represent and take part in social interactions: Comparing individuals with high-functioning autism and typically developing controls. *The European Journal of Developmental Psychology, 7*(1), 104–122. http://doi.org/10.1080/17405620903024263

Berger, N. I., & Ingersoll, B. (2014). A further investigation of goal-directed intention understanding in young children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *44*(12), 3204–3214. http://doi.org/10.1007/s10803-014-2181-z

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*, *19*(3), 354–369. http://doi.org/10.1177/1073191112446655

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*(5), 382–386. http://doi.org/10.1111/j.1467-9280.2007.01909.x

Biro, S., & Leslie, A. M. (2007). Infants' perception of goal-directed actions: development through cue-based bootstrapping. *Developmental Scienc*e, *10*(3), 379-398.

Bottiroli, S., Cavallini, E., Ceccato, I., Vecchi, T., & Lecce, S. (2016). Theory of Mind in aging: Comparing cognitive and affective components in the faux pas test. *Archives of Gerontology and Geriatrics*, *62*, 152–162.

Bowler, D. M. (1992). "Theory of Mind" in Asperger's Syndrome Dermot M. Bowler. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *33*(5), 877–893. http://doi.org/10.1111/j.1469-7610.1992.tb01962.x

Broekhof, E., Ketelaar, L., Stockmann, L., van Zijp, A., Bos, M. G. N., & Rieffe, C. (2015). The Understanding of Intentions, Desires and Beliefs in Young Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *45*(7), 2035–2045. http://doi.org/10.1007/s10803-015-2363-3

Buon, M., Dupoux, E., Jacob, P., Chaste, P., Leboyer, M., & Zalla, T. (2013). The role of causal and intentional judgments in moral reasoning in individuals with high functioning autism.

Journal of Autism and Developmental Disorders, *43*(2), 458–470. http://doi.org/10.1007/s10803-012-1588-7

Bromberger, S. (1970). Why-questions. In B. Brody (ed)., *Readings in the Philosophy of Science.* Englewood Cliffs, NJ: Prentice-Hall.

Buss, A. R. (1978). Causes and reasons in attribution theory: A conceptual critique. *Journal of Personality and Social Psychology, 36,* 1311–1321. doi:10.1037/0022-3514.36.11.1311

Carpenter, M., Pennington, B. F., & Rogers, S. J. (2001). Understanding of others' intentions in children with autism. *Journal of Autism and Developmental Disorders*, *31*(6), 589–599.

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain: A Journal of Neurology*, *125*(8), 1839–1849.

Channon, S., Lagnado, D., Fitzpatrick, S., Drury, H., & Taylor, I. (2011). Judgments of cause and blame: sensitivity to intentionality in Asperger's syndrome. *Journal of Autism and Developmental Disorders*, *41*(11), 1534–1542. http://doi.org/10.1007/s10803-011-1180-6

Clement, R. W., & Krueger, J. (2000). The primacy of self-referent information in perceptions of social consensus. *British Journal of Social Psychology, 39,* 279–299. doi:10.1348/014466600164471

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21. http://doi.org/10.1016/j.cognition.2012.11.008

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108,* 353–380. doi:10.1016/j.cognition.2008.03.006

Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy, 60*(23), 685–

700.

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences, 1*(04), 568–570. http://doi.org/10.1017/S0140525X00076664

Dennett, D. C. (1987). *The intentional stance.* Cambridge: MIT Press.

Devine, P. G., Sedikides, C., & Fuhrman, R.W. (1989). Goals in social information processing: The case of anticipated interaction. *Journal of Personality and Social Psychology, 56*, 680-690.

DiGiovanni, M., Korman, J. & Malle, B. F. (2016, August). How we explain puzzling behavior: The differential use of beliefs and desires. Poster presented to the Brown Summer Research Symposium.

Dretske, F. (1988/1992). *Explaining behavior: Reasons in a World of Causes*. Cambridge: MIT Press.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*, 327–339. doi:10.1037/0022-3514.87.3.327

Fadda, R., Parisi, M., Ferretti, L., Saba, G., Foscoliano, M., Salvago, A., & Doneddu, G. (2016). Exploring the Role of Theory of Mind in Moral Judgment: The Case of Children with Autism Spectrum Disorder. *Frontiers in Psychology, 7*, 523.

Feinfield, K. A., Lee, P. P., Flavell, E. R., Green, F. L., & Flavell, J. H. (1999). Young children's understanding of intention. *Cognitive Development, 14*(3), 463–486.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York, NY: McGraw-Hill.

Frith, U., Morton, J., & Leslie, A. M. (1991). The cognitive basis of a biological disorder: autism. *Trends in Neurosciences, 14*(10), 433–438.

Gergely, G., & Csibra, G. (1997). Teleological reasoning in infancy: the infant's naive theory of rational action. A reply to Premack and Premack. *Cognition*, *63*(2), 227–233.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naıve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

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

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*, 21–38. doi:10.1037/0033-2909.117.1.21

Gillberg, C., Gillberg, C., Råstam, M., & Wentz, E. (2001). The Asperger Syndrome (and high-functioning autism) Diagnostic Interview (ASDI): a preliminary study of a new structured clinical interview. *Autism: The International Journal of Research and Practice*, *5*(1), 57–66. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11708390

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371–395.

Haigh, M., & Bonnefon, J.-F. (2015). Eye Movements Reveal How Readers Infer Intentions From the Beliefs and Desires of Others. *Experimental Psychology*, *62*(3), 206–213.

Hamilton, A. F. de C. (2009). Research review: Goals, intentions and mental states: challenges for theories of autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *50*(8), 881–892. http://doi.org/10.1111/j.1469-7610.2009.02098.x

Happé, F. G. (1994). An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154.

Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66*(3), 843–855.

Harman, G. H. (1965). The Inference to the Best Explanation. *The Philosophical Review*, *74*(1), 88–95.

Hassin, R. R., Aarts, H., & Ferguson, M. J. (2005). Automatic goal inferences. *Journal of Experimental Social Psychology, 41*, 129–140. doi:10.1016/j.jesp.2004.06.008

Hastorf, A. H., Schneider, D. J., & Polefka, J. (1970). *Person perception*. Reading, MA: Addison-Wesley.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Hempel, C. G. (1966). *Philosophy of natural science*. Englewood Cliffs, NJ: Prentice-Hall,.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*, 65–81.

Hilton, D. J. (2007). Causal explanation: From social perception to knowledge-based causal attribution. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 232–253). New York, NY: Guilford Press.

Hilton, D. J., Smith, R. H., & Kim, S. H. (1995). Processes of causal explanation and dispositional attribution. *Journal of Personality and Social Psychology, 68*, 377–387. doi:10.1037/0022-3514.68.3.377

Hobson, R. P., & Lee, A. (1999). Imitation and identification in autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *40*(4), 649–659.

Hogrefe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance versus False Belief: A Developmental Lag in Attribution of Epistemic States. *Child Development*, *57*(3), 567–582.

Johnson, J. T., Jemmott, J. B., & Pettigrew, T. F. (1984). Causal attribution and dispositional inference: Evidence of inconsistent judgments. *Journal of Experimental Social Psychology*, *20*, 567–585. doi:10.1016/0022-1031(84)90044-1

Jolliffe, T., & Baron-Cohen, S. (1999). The Strange Stories Test: a replication with high-functioning adults with autism or Asperger syndrome. *Journal of Autism and Developmental Disorders*, *29*(5), 395–406. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10587886

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219–266). New York: Academic Press.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*, 1–24. doi:10.1016/0022-1031(67)90034-0

Kalish, C. W., & Shiverick, S. M. (2004). Children's reasoning about norms and traits as motives for behavior. *Cognitive Development*, *19*, 401–416.

Karniol, R. (2003). Egocentrism versus protocentrism: the status of self in social prediction. *Psychological Review*, *110*(3), 564–580.

Kashima, Y., McKintyre, A., & Clifford, P. (1998). The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology*, *1*(3), 289–313.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 15, pp. 192–240). Lincoln: University of Nebraska Press.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific Explanation* (Vol. 8, pp. 410–505). Minneapolis: University of Minnesota Press.

Keil, F.C. (2006). Explanation and Understanding. *Annual Review of Psychology.* 57, 227-254.

Klin, A. (2000). Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: the Social Attribution Task. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *41*(7), 831–846. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/1469-7610.00671/abstract

Klin, A., Jones, W., Schultz, R., & Volkmar, F. (2003). The enactive mind, or from actions to cognition: lessons from autism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *358*(1430), 345–360. http://doi.org/10.1098/rstb.2002.1202

Knobe, J. (2003). "Intentional Action and Side Effects in Ordinary Language." *Analysis*, 63, 190-193.

Korman, J., Cusimano, C. Monroe, A., Smith, J., & Malle, B. F. (2014, July). Not so bad after all? The role of explanation features in blame mitigation. Poster presented at the Annual Meeting of the Cognitive Science Society, Quebec City, Canada.

Korman, J., & Malle, B. F. (2016). The folk concept of rationality. Unpublished Data, Brown University.

Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: neuropsychological evidence from autism. *Cognition*, *43*(3), 225–251. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1643814

Lewis, P. T. (1995). A naturalistic test of two fundamental propositions: Correspondence bias and the actor-observer hypothesis. *Journal of Personality*, *63*, 87–111.

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*, 551–556. doi:10.1016/j.jesp.2009.12.019

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?" *Cognition, 110*, 248–253. doi:10.1016/j.cognition.2008.10.007

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., … Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. *Journal of Autism and Developmental Disorders*, *30*(3), 205–223. http://doi.org/10.1023/A:1005592401947

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *24*(5), 659–685. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7814313

Loth, E., Gómez, J. C., & Happé, F. (2008). Event schemas in autism spectrum disorders: the role of theory of mind and weak central coherence. *Journal of Autism and Developmental Disorders*, *38*(3), 449–463. http://doi.org/10.1007/s10803-007-0412-2

Loth, E., Gómez, J. C., & Happé, F. (2011). Do high-functioning people with autism spectrum disorder spontaneously use event knowledge to selectively attend to and remember context-relevant aspects in scenes? *Journal of Autism and Developmental Disorders 41*(7), 945-61.

MacFarquhar, L. (2015, August 3). The children of strangers. *The New Yorker*.

Malle, B. F. (1998). *F.Ex: Coding scheme for people's folk explanations of behavior*. Latest version 4.5.7 (2014), Retrieved August 2015 from http://research.clps.brown.edu/SocCogSci/CodingSchemes.html.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*, 23–48. doi:10.1207/s15327957pspr0301_2

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, *132*, 895–919. doi:10.1037/0033-2909.132.6.895

Malle, B. F. (2011). Time to give up the dogmas of attribution: A new theory of behavior explanation. In M. P. Zanna & J. M. Olson (Eds.), *Advances of Experimental Social Psychology* (Vol. 44, pp. 297–352). San Diego, CA: Academic Press.

Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology*, *102*, 661–684. doi:10.1037/a0026790

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A Theory of Blame. *Psychological Inquiry*, *25*(2), 147–186. http://doi.org/10.1080/1047840X.2014.877340

Malle, B. F., & Knobe, J. (1997). The Folk Concept of Intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101–121.

Malle, B. F., & Knobe, J. (1997). Which behaviors do people explain? A basic actor-observer asymmetry. *Journal of Personality and Social Psychology*, *72*, 288–304. doi:10.1037/0022-3514.72.2.288

Malle, B. F., Knobe, J., & Nelson, S. E. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology*, *93*, 491–514. doi:10.1037/0022-3514.93.4.491

Malle, B. F., Knobe, J., O'Laughlin, M. J., Pearce, G. E., & Nelson, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person-situation attributions. *Journal of Personality and Social Psychology, 79,* 309–326. doi:10.1037/0022-3514.79.3.309

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA, US: MIT Press.

Malle, B. F. (2011). Time to give up the dogmas of attribution: An alternative theory of behavior explanation. In: Olson J. M., Zanna M. P., (Eds), *Advances in experimental social psychology,* Vol. 44. (pp. 297-352). Burlington, VT: Academic Press.

McClure, J. (2002). Goal-based explanations of actions and outcomes. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 12, pp. 201–235). New York: Wiley.

McClure, J., & Hilton, D. J. (1998). Are goals or preconditions better explanations? It depends on the question. *European Journal of Social Psychology, 28,* 897–911.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*(3), 440–466. Meltzoff, A. N. (1995). Understanding the Intentions of Others: Re-Enactment of Intended Acts by 18-Month-Old Children. *Developmental Psychology, 31*(5), 838–850. http://doi.org/10.1037/0012-1649.31.5.838

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences*, *108*(7), 2688–2692. http://doi.org/10.1073/pnas.1011734108

Moskowitz, G. B., & Olcaysoy Okten, I. (2016). Spontaneous goal inference (SGI). *Social and Personality Psychology Compass, 9*, 644–661.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. New York, NY: Oxford University Press.

O'hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution / British Ecological Society, 1,* 118-122. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2010.00021.x/full

O'Laughlin, M. J., & Malle, B. F. (2002). How people explain actions performed by groups and individuals. *Journal of Personality and Social Psychology*, *82*, 33–48. doi:10.1037/0022-3514.82.1.33

Peterson, C. C., Wellman, H. M., & Liu, D. (2005). Steps in Theory-of-Mind Development for Children With Deafness or Autism. *Child Development*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8624.2005.00859.x/full

Peterson, D. M., & Bowler, D. M. (2000). Counterfactual Reasoning and False Belief Understanding in Children with Autism. *Autism: The International Journal of Research and Practice*, *4*(4), 391–405. http://doi.org/10.1177/1362361300004004005

Phillips, W., Baron-Cohen, S., & Rutter, M. (1998). Understanding intention in normal development and in autism. *The British Journal of Developmental Psychology*, *16*(3), 337–348.

Premack, D. (2010). Why Humans Are Unique: Three Theories. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *5*(1), 22–32.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, *1*(04), 515–526.

Ratcliffe, M. (2007). *Rethinking commonsense psychology: A critique of folk psychology, theory of mind and simulation*. New York, NY: Palgrave Macmillan.

Raven, J., Raven, J. C., & Court, J. H. (1998c). Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3, The Standard Progressive Matrices. Oxford, England: Oxford Psychologists Press/San Antonio, TX: The Psychological Corporation.

Read, S. J., Jones, D. K., & Miller, L. C. (1990). Traits as goal-based categories: The importance of goals in the coherence of dispositional categories. *Journal of Personality and Social Psychology, 58,* 1048–1061. doi:10.1037/0022-3514.58.6.1048

Reeder, G. D. (2009). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological Inquiry, 20,* 1–18. doi:10.1080/10478400802615744

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86,* 61–79. doi:10.1037/0033-295X.86.1.61

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14- and 18-month-olds. *Developmental Psychology, 33*(1), 12–21.

Richardson, H.L., Baker, C.L., Tenenbaum, J.B., & Saxe, R.R. (2012). The Development of Joint Belief-Desire Inferences. In *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society* (pp. 923-928).

Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998/1). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development, 13*(1), 73–90. http://doi.org/10.1016/S0885-2014(98)90021-1

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (vol. 10). New York: Academic Press.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation*. New York: McGraw-Hill.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sanbonmatsu, D. M., Mazur, D., Behrends, A. A., & Moore, S. M. (2015). The role of the baserate frequency of correspondent behavior and trait stereotypes in attribution: Building on Rothbart and Park (1986). *Social Cognition, 33,* 255–283. doi:10.1521/soco.2015.33.4.255

Sandis, C. (Ed.). (2009). *New essays on the explanation of action*. New York, NY: Palgrave Macmillan.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 87–124. doi:10.1146/annurev.psych.55.090902.142044

Schachner, A. & Carey, S. (2013). Reasoning about 'irrational' actions: When intentional movements cannot be explained, the movements themselves are seen as the goal. *Cognition*, 129, 309-327.

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

Schult, C. A. (1996). Intended actions and intentional states: Young children's understanding of the causes of human actions. Doctoral Dissertation, University of Michigan.

Schult, C. A. (2002). Children's understanding of the distinction between intentions and desires. *Child Development*, *73*(6), 1727–1747.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press. Retrieved from https://books.google.com/books?id=nAYGcftgT20C

Shaver, K. G. (1975). *An introduction to attribution processes*. Cambridge, MA: Winthrop.

Shulman, C., Guberman, A., Shiling, N., & Bauminger, N. (2012). Moral and social reasoning in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *42*(7), 1364–1376.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*, 131–142. doi:10.1037/0033-2909.105.1.131

Slugoski, B. R., Lalljee, M., Lamb, R., & Ginsburg, G. P. (1993). Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology*, *23*, 219–238. doi:10.1002/ejsp.2420230302

Snodgrass, S. R. (1976). The development of trait inference. *Journal of Genetic Psychology*, *128*, 163–172.

Solomon, A. (2014, March 17). The reckoning. *The New Yorker*.

Tager-Flusberg, H., & Joseph, R.M. (2005). How language facilitates the acquisition of false belief understanding in children with autism. In J. Astington & J. Baird (Eds.), Why language matters for theory of mind (pp. 298–318). Oxford, UK: Oxford University Press.

Tager-Flusberg, H., & Sullivan, K. (1995). Attributing mental states to story characters: A comparison of narratives produced by autistic and mentally retarded individuals. *Applied Psycholinguistics*, *16*(03), 241–256. http://doi.org/10.1017/S0142716400007281

Tan, J., & Harris, P. L. (1991). Autistic children understand seeing and wanting. *Development and Psychopathology*, *3*(02), 163–174. http://doi.org/10.1017/S0954579400000055

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *The Behavioral and Brain Sciences*, *28*(5), 675–91; discussion 691–735. http://doi.org/10.1017/S0140525X05000129

Trafimow, D., Bromgard, I.K., Finlay, K.A., & Ketelaar, T. (2005). The role of affect in determining the attributional weight of immoral behaviors. *Personality and Social Psychology Bulletin, 31*, 935-948.

Trillingsgaard, A. (1999). The script model in relation to autism. *European Child & Adolescent Psychiatry*, *8*(1), 45–49.

Uleman, J. S., Saribay, S. A., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, *59*, 329–360. doi:10.1146/annurev.psych.59.103006.093707

Vallacher, R. R., & Wegner, D. M. (1987). What do people think they're doing? Action identification and human behavior. *Psychological Review*, *94*(1), 3.

Van Overwalle, F., Van Duynslaeger, M., Coomans, D., & Timmermans, B. (2012). Spontaneous goal inferences are often inferred faster than spontaneous trait inferences. *Journal of Experimental Social Psychology*, *48*, 13–18. doi:10.1016/j.jesp.2011.06.016

Verschoor, S., & Biro, S. (2012). The Primacy of Means Selection Information Over Outcome Selection Information in Infants' Goal Attribution. *Cognitive Science*, *36*, 714–725.

Vivanti, G., Nadig, A., Ozonoff, S., & Rogers, S. J. (11/2008). What do children with autism attend to during imitation tasks? *Journal of Experimental Child Psychology*, *101*(3), 186–205. http://doi.org/10.1016/j.jecp.2008.04.008

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*, 343–357. doi:10.1016/j.cognition.2014.07.008

Wechsler, D., Coalson, D. L., & Raiford, S. E. (2008). WAIS-IV: Technical and Interpretive Manual. San Antonio, TX: Pearson.

Weiner, B. (1985). "Spontaneous" causal thinking. *Psychological Bulletin, 97*, 74-87.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. *Child Development*, 72, 655–684.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: the early development of everyday psychology. *Cognition*, *35*(3), 245–275.

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.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*, 776–806. doi:10.1111/j.1551-6709.2010.01113.x

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology, 47*, 237–252. doi:10.1037/0022-3514.47.2.237

Wong, P.T.P. & Weiner, B. (1981). When people ask "why" questions and the heuristics of attributional search. *Journal of Personality and Social Psychology, 40*, 650-663.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America, 104*(20), 8235–8240. http://doi.org/10.1073/pnas.0701408104

Zalla, T., Barlassina, L., Buon, M., & Leboyer, M. (2011). Moral judgment in adults with autism spectrum disorders. *Cognition, 121*(1), 115–126. http://doi.org/10.1016/j.cognition.2011.06.004

Zalla, T., & Leboyer, M. (2011). Judgment of Intentionality and Moral Evaluation in Individuals with High Functioning Autism. *Review of Philosophy and Psychology, 2*(4), 681–698. http://doi.org/10.1007/s13164-011-0048-1

Zalla, T., Sav, A.-M., Stopin, A., Ahade, S., & Leboyer, M. (2009). Faux pas detection and intentional action in Asperger Syndrome. A replication on a French sample. *Journal of Autism and Developmental Disorders, 39*(2), 373–382.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162–185.

APPENDIX

Instructions: Chapter 2 studies 1, 2, & 4 and Chapter 3

Main Instructions

In the following task, you will read a series of sentences.

In the space provided, please add whatever sentences or phrases you think are needed to make sense of the sentence. Please do not negate any part of the sentence as it is, but rather, only *add* information in the box provided to help the original sentence make better sense.

In other words, try to imagine a world in which the sentence is true, and then add information onto the sentence to help it make better sense.

For some sentences, adding a short phrase may be sufficient. For others, you may feel that adding a couple of sentences is required to make sense of the original sentence.

Please give each sentence your best effort. When your response is complete, you should be satisfied that the information you added makes sense of the sentence given.

[In Chapter 2, studies 1 & 2, the following paragraph appeared in the order here. In Chapter 2, Study 4 and Chapter 3, participants completed one practice trial (identical in format to the main study trials) before beginning of the task, and the following paragraph was added after participants had given their open-ended response to the sentence in the practice trial.]

After you have typed your response, you will be asked to evaluate it. Taking into account the information you added, how much SENSE the situation described in the original sentence make NOW? Here you will rate your added information on a scale from 1 (No sense at all) to 8 (Perfect sense).

Instructions for each trial

For each trial, the participant is presented with the stimulus sentence. They receive the following instruction:

In the box below, add whatever information you think is needed for the sentence above to make better sense.

After pressing a "continue" button, they are taken to the next screen.

Here is the information you added: [Participant's added information from previous screen]

Having added your information, how much SENSE does the situation described in the original sentence make NOW?
[Rated from 1: No sense at all to 8: Perfect sense]

Note: Throughout the task, instructions related to the stimulus (including the stimulus itself) are highlighted in *green*, while instructions related to the participant's response are highlighted in *yellow*.

Instructions: Chapter 2, Study 3

Main Instructions:
[Unless otherwise noted, instructions were presented on the computer.]

In each trial you will be presented with a sentence. Once the sentence has appeared, read it out loud.

Then you will be asked to ADD WHATEVER INFORMATION YOU THINK IS NEEDED to MAKE SENSE of the sentence.

Please SPEAK your responses OUT LOUD into the microphone.

When your response is complete, you should be SATISFIED that the information you added MAKES SENSE of the sentence given.

Once the ADD INFORMATION prompt appears, you will have 45 seconds to complete your response. You will receive a warning when there are 15 seconds remaining.

In your response, please DO NOT CHANGE details of the sentence. ALL parts of the sentence, as WELL as the meaning of the sentence AS A WHOLE, should remain true even AFTER your information is added. That is, try to add information that will make the ENTIRE sentence AS A WHOLE easier to understand."

For some sentences, adding a short phrase may be sufficient. For others, you may feel that adding a couple of sentences is required to make sense of the sentence. After you have spoken your response, you will then be asked to rate how well your response makes sense of the sentence given.

Please speak loudly and clearly, and in complete phrases. The first two sentences are practice sentences. Let the experimenter know when you are ready to begin the practice sentences.

[Experimenter re-enters the room and participant completes practice sentences. The experimenter then reminds participants: "Make sure not to change or negate any details of the stimulus sentence. You should only add information to help make the sentence make better

sense. What you want to do is to imagine a world in which the sentence is true." Before leaving the room, the experimenter then answers any final questions the participant may have.]

The practice trials have completed. Press the space bar when you are ready to begin the experiment.

Instructions for each Trial

For each trial, participants saw the following:

Please READ the following sentence out loud.

[Sentence is presented]

Now please speak your response. Add whatever SENTENCES or PHRASES you think are needed to MAKE SENSE of the above sentence. Remember to only ADD ADDITIONAL information. Do not CHANGE the details of any part of the sentence.

You have 15 seconds remaining to finish up your response. When you are finished, hit return."

With the information you added, the situation described in the sentence above should now make better sense.

Having added your information, how much SENSE do you think the sentence NOW makes?

Press the number at the top of the keyboard corresponding to your response. (1 – No sense at all; 5 – Some Sense; 9 – Perfect Sense).

Chapter 2 & 3: Stimuli

Chapter 2, Study 1 and 2 Stimuli

Category 1: Schema-breaking, Script-compliant

1. He started the car with a hairbrush.
2. The chef chopped up the cassette tapes using his sharpest knife.
3. He brought a tape recorder on a first date.
4. The janitor cleaned the floors with honey.*

Category 2: Script-breaking, Schema-compliant

1. The garbage men dropped off bags of trash at the end of each driveway.
2. The woman emptied a bucket of dirt into her washing machine.
3. The zookeeper played a concerto for the animals.
4. The man burnt CDs for his cat.

Category 3: Script-breaking, Schema-breaking

1. The chief executive taught yoga in the town junkyard.
2. The teenager gave her typewriter a bath under the bridge.
3. The supermarket owner painted her scarecrow magenta.
4. The priest served his coffee cup in tomato sauce.

Category 4: Script-compliant, Schema-compliant

1. The vacationers brought six hundred cases of beer to the beach.
2. The young woman drank eight gallons of Gatorade before the race.
3. The folk singer gave each of his fans three hundred copies of his latest album.
4. The little league soccer team practiced for 35 hours straight.

*In Study 2, this item replaced an unusable item from Study 1.

Chapter 2, Study 3 Stimuli

Category 1: Schema-breaking, script compliant

Marked* desires.

1. She made some chocolate cake because she wanted her brother to develop an appreciation for art.
2. The burly hikers applied a large amount of sunscreen because they wanted to ward off the bugs.
3. He ran for president because he wanted his father to be cured of cancer.
4. The women of the knitting group met because they wanted to include more cats.

Unmarked desires.

1. She went to the plant nursery to pick up some video games for her son.
2. The man went to the toy store so he could buy some watermelons.
3. She went to the eye doctor so she could get her muscle pains checked out.
4. The tax collector knocked on the family's front door to get some chocolate.

Marked beliefs.

1. He brought a blank piece of white paper to the birthday party because he thought it would make an excellent gift.
2. The woman tried on the scarf because she thought it would cut off her circulation.
3. The man decided to get his leg amputated because he thought the scab on his knee was ugly.
4. The three sisters played games in the yard because they thought their mother would bring them a pony.

Unmarked beliefs.

1. The lawyer called the police because his client was wearing an ugly suit.
2. The woman put on her oven mitts because it was time to do the dishes.
3. The man bought a diamond ring for his wife because they were celebrating her mother's wedding anniversary.
4. The 15-year-old begged her mother for the new encyclopedia set because it would make her more popular at school.

* "Marked" and "unmarked" refer to the presence or absence of a mental state marker verb (e.g., "he thought," "she wanted").

Category 2: Script-breaking, schema-compliant

Marked desires.

1. The boys went to the rock concert because they wanted to drown out all the loud music.
2. The flight attendant walked up and down the aisle because she wanted to see what the passengers were wearing.
3. The committee chairman voted to install the new cell phone tower in the forest because he wanted to protect the environment.
4. The preschool teacher went to get the toys from the closet because she planned to throw them at the three-year-olds.

Unmarked desires.

1. She got out her keys to carve a note on the door for her sister.
2. The garbage men drove their truck all around the city to drop off a can full of trash at the end of each driveway.
3. The salesman handed his customer a dress so she could tie it around her neck.
4. The woman dressed up her pet poodle so she could take it for a walk in the stroller.

Marked beliefs.

1. She invited her family to her wedding because she knew they would throw dresses at her.
2. He bought all the ingredients to make a stuffed turkey because he knew his aunt was away in Europe for the year.
3. Grandma began to bake some bread because she knew the children would play with the oven.
4. The man drove his golf cart rapidly around the parking lot because he thought the golf clubs would all fall out.

Unmarked beliefs.

1. The woman called the carpenter because he could help with her collection of hammers.
2. The college graduate decided to study investment banking because it would lose him a lot of money.

3. She went to church because the basketball team was triumphing over its most hated rival.
4. Grandpa began to make a big meal because his family was coming with an enormous amount of delicious food.

Category 3: Schema-breaking, script breaking

Marked desires.

1. The journalist shot photos of the crime scene because he wished the weather would improve.
2. The accountant got his books in order because he wanted the ice cream truck to pass on the street.
3. She put on a pair of flip flops because she wanted to improve her cattle herding skills.
4. She went to the welfare office because she wanted to buy a diamond watch.

Unmarked desires.

1. The young man purchased an expensive oriental rug to improve his vision.
2. The woman called the taxi so she wouldn't be overtaken by the ocean waves.
3. The women watched the space shuttle take off so that the restaurant would stay open for a few more hours.
4. He offered to buy the girl a drink so that she would think that there were aliens landing on earth that day.

Marked beliefs.

1. He jumped off the building because he thought it was too cold out for iced coffee.
2. The woman went to get the mail because she thought she had a very serious illness.
3. She picked up the carton of milk because she thought lamp would turn on.
4. He flew the airplane into the tree because he knew it was dinnertime soon.

Unmarked beliefs.

1. He put on his headphones because a giraffe appeared outside.
2. She served orange juice to the guests because the roof was leaking.
3. The manager closed his store because the trees were changing colors.
4. She got her dress hemmed because the faucet was dripping.

Category 4: Sufficiency items

Marked desires.

1. The customer left a 5 cent tip because he wanted to show his appreciation for the restaurant's excellent service.
2. The nurse waited 3 years before reporting the patient's death because she wanted to be certain he was really dead.
3. The woman wrote a letter to the editor because she hoped to win a Pulitzer Prize.
4. He played the lottery twice because he wanted to ensure he would win.

Unmarked desires.

1. The fisherman caught eight giant fish so his wife could put dinner on the table.
2. She planted a garden in their backyard to feed the country's armed forces.
3. He poured four gallons of laundry detergent into the washing machine so his clothes would come out clean.
4. The woman got out her watering can to put out the fire in the kitchen.

Marked beliefs.

1. He took 35 sleeping pills because he knew it would give him a good night's sleep.
2. She drank five gallons of Gatorade because she knew it would hydrate her before the race.
3. He read the entire dictionary cover to cover because he thought it would increase his vocabulary.
4. The woman spent her life savings on 5,000 boxes of Girl Scout cookies because she knew the neighbor's daughter was selling them.

Unmarked beliefs.

5. The teenager stole everything in the electronics store because his radio had run out of batteries.
6. The woman screamed for half an hour because she had missed her train.
7. The man rented three semi trucks because he was transporting his dining room set to his new apartment.
8. She chopped down the forest because it gave her some wood for her camp fire.

Chapter 2, Study 4 Stimuli

Category 1: Incompatible Causal History Only

1. The man went to the toy store because he's a gun enthusiast.
2. The journalist shot photos of the crime scene because he often puts them on his nightstand.
3. She picked up the carton of milk because she was at the fanciest restaurant in town.
4. He put on his headphones because he had always loved German expressionism.

Category 2: Incompatible Reason with Causal History*

Category 2a: Researcher-generated Reason + Causal History Pairs.

1. The flight attendant walked up and down the aisle because she wanted to see what the passengers were wearing; she was an introvert.
2. The man bought a diamond ring for his wife because they were celebrating his mother-in-law's wedding anniversary; they had a family tradition.
3. The young man purchased an expensive oriental rug to improve his vision; he had a lot of money to waste.
4. The tax collector knocked on the family's front door to get some chocolate; he knew the family.

Category 2b: Participant-generated Reason + Causal History Pairs.

1. He ran for president because he wanted his mother to be cured of cancer; he was a doctor.
2. Grandma began to bake some bread because she knew the children would play with the oven; she had a close relationship with them.
3. She went to the plant nursery to pick up some video games for her son; she was an indulgent parent.
4. Grandpa began to make a big meal because his family was coming with an enormous amount of delicious food; his family members were all women.

*Between subjects, each item in this category is presented in both orders: Reason first and Causal History first (only Reason first is shown here).

Chapter 3: Stimuli

Goal Blocking - Belief

The film director selected an Oscar-winning actor for the lead role because he thought that the actor had no talent.

The chef baked peanuts into the bride's dessert because he knew she had a serious peanut allergy.

Goal Blocking – Desire

The middle-aged man hung up on his sister because he wanted her to know how much he loved her.

The CEO scolded his office staff so that they would throw themselves a party.

Means-End: Belief

She strapped on her cross-country skis because she knew the supermarket was about to close.

He picked up his frying pan because it was time to put together the Ikea furniture.

Means-End: Desire

The woman sent her friend a check for $1,000 because she wanted to pay her friend back for lunch.

The lawyer placed a call to one of his clients because he wanted to see what the weather was like.

Chapter 4: Measures

Control Stories

False Belief, Positive

Thomas is a wealthy businessman who is well known in the philanthropic community for his frequent charitable donations. One day Thomas gives a large anonymous donation to an organization that provides clean water in third world countries. A few weeks later, at a business meeting he attends, he sees a charity appeal by an AIDS education organization, but Thomas abstains from giving. Nonetheless, the AIDS education fund drive is very successful. The AIDS charity president sees Thomas in the hotel lobby and says to him, “Thanks to your generosity, people in the third world will be lifted up out of poverty and disease.”

Juan, a long time art lover, frequently goes to galleries to view art being produced by up-and-coming artists. Juan is shy and so rarely meets or talks to the artists. At one showing he sees a painting he thinks is quite beautiful. Sitting down next to him is the artist, who is acting quiet because so far his art has not garnered much attention. Juan is so enthusiastic about the piece that he has to tell someone, so he turns to the man next to him and says to him, "This is quite a beautiful painting, isn’t it?"

True Belief, Negative

Elorie is an avid skier and makes a point to go skiing as many times as she can during the winter season. On one weekend at the resort, her leg feels sore, so she decides to take it easy on the slopes. As she is skiing down the mountain, a snowboarder flies by her at a very high speed, inches away from crashing into her. When she gets to the bottom of the mountain, she confronts him, saying, "You jerk! You almost got me killed!"

Ethan is a scientist who studies the biology of certain bugs and butterflies. He has accumulated several thousand different species in his laboratory. One day his research assistant, Steven, accidentally spills a very corrosive acid on several of the rarer bug samples. When Ethan comes into the lab he immediately notices the damage. He finds Steven and yells, "You are a worthless assistant!"

True Belief, Positive

Patricia is a well-respected and highly demanding food critic for an elite culinary magazine. Many of her most recent restaurant trips have been very disappointing. The next restaurant on her list to review has just opened and is being run by a talented young chef. When she goes there, she orders a simple entrée of duck. She is not expecting much and so is completely surprised by how excellent it tastes. When she sees the chef on her way out of the restaurant, she goes up to him and says, "This is the best restaurant I have been to in a long time."

Paul’s phone has been giving him a lot of trouble. He takes it in to the store to have someone look at it. Stephanie works there and runs a test that reveals a major bug in the phone’s software. She replaces Paul’s phone with a working model. Later in the week Paul is walking by the phone store and notices that she is working again. He runs inside and says to her, "Thank you again for all your help earlier this week. You were amazing!"

Control Stories (Modification in Study 7)

False Belief, Positive

Thomas is a wealthy businessman who is well known in the philanthropic community for his frequent charitable donations. One day Thomas gives a large anonymous donation to research on breast cancer in the United States. A few weeks later, at a business meeting he attends, he sees a charity appeal by an international AIDS education organization, but Thomas abstains from giving. Nonetheless, the AIDS education fund drive is very successful. The AIDS charity president sees Thomas in the hotel lobby and says to him, "Thanks to your generosity, people in the third world will be free of this terrible disease."

Faux Pas Stories

(Added information in 3 conditions presented in italics)

School

Clara is very short and dresses plainly. One day she goes to pick up her son James from school early for a medical appointment. Clara enters the school and spots James's teacher, Mrs. Hayes. Mrs. Hayes thinks that *Clara is a student lost in the hallway*. (B) *Mrs. Hayes wants to help* (D). Before Clara can ask after James' whereabouts, Mrs. Hayes looks at Clara and says, "Have you lost your class, honey?"

Squash

Sean and James are squash buddies who live in a very politically liberal city on the west coast of the US. Sean, however, is a devoted evangelical Christian and staunch Republican, and is particularly passionate in his opposition to abortion. Although Sean and James often talk about squash and other sports, the subject of politics has never come up. *James thinks Sean must be very liberal like him.* (B) One day, James's wife comes to pick him up from his squash game. *James wants Sean to join him in supporting his political causes.* (D) James says, "Sean, I'd like you to meet my wife. Together, she and I are on a mission to rid the world of anti-choice Republicans!"

Breakup

Chloe has just gone through a very bad breakup with her boyfriend Jake. Since the breakup, she has become very depressed, eating doughnuts and Twinkies by the boxful, and she has begun to gain some weight. Chloe's friend, Liana, has spent the year abroad, and *thinks that Chloe and Jake are still together.* [B] and when she arrives back in the U.S. Chloe goes to pick her up from the airport. *Liana really wants to hear all the latest news about what happened back home.* [D] As soon as Liana sees Chloe, she exclaims, "Oh my GOSH! This is so exciting! Are you and Jake pregnant?!"

Gas station

For his whole life, Sandeep has lived in New Jersey, a state in the USA where it is illegal to pump one's own gas. Sandeep has just moved to Pennsylvania. When he arrives at his

neighborhood gas station, he does his usual routine, parking in front of one of the pumps and waiting for the attendant. When no one comes to serve him, he gets out of the car and looks around. He sees an older man pumping gas into an SUV at the adjacent pump, and *he thinks that the man works for the gas station.* [B] *He wants to get his gas pumped*, so he says, [D] "When you're done over there, I'd like $20 worth of regular, please."

Twins
John has recently been laid off from his job as an assembly line worker in a local factory and is feeling sorry for himself. John has an identical twin brother, George, who is a wealthy businessman. John gets dressed up one day to attend a family wedding. While waiting for the ceremony to start, he begins chatting with a group of other guests seated nearby. One woman turns to John and begins a conversation**.** *She thinks that John is actually George.* [B] *She wants to express her concern for his family,* [D**]** so she says, "I'm so sorry to hear that your brother just got fired. It must be so hard to see your brother's long line of failures."

Resort
Claudia and Wayne are a married couple. She is 25 years his junior. They travel to Costa Rica together for vacation. When they arrive at the resort with their luggage, they are greeted by the bell hop. *The bell hop thinks that Claudia and Wayne are father and daughter.* [B] *Wanting to make them feel welcome* [D], he remarks, "What a nice surprise to see a father and daughter coming to enjoy the resort."

Family Reunion
While attending a family reunion, Su Lin and her four siblings decide to stay in the same hotel. As she heads downstairs to meet her family for dinner, Su Lin trips and falls down the stairs, taking a hard fall onto her face. Although she is not seriously injured, her face is cut and bruised. Su Lin would be terribly embarrassed if anyone found out, so she returns to her room and spends an hour applying powder and cream to conceal the marks. Su Lin's brother *Chin Ho thinks she didn't put effort into her appearance.* [B] *He wants to make sure that everyone looks their best for the big reunion dinner.* [D] When he sees Su Lin, Chin Ho remarks, "Su! Have you thought about putting on some makeup for this special occasion?"

Vance and Leonard are partners at a law firm. Both men are happily married but often have to spend long hours at the office together working on cases. One day Vance and Leonard are walking down the hallway discussing new business with a client. One of the law firm's newly hired secretaries, Martha, is very chatty. Because she sees them together often, she *thinks that Vance and Leonard are a gay couple.* [B] *To make conversation,* [D]when she walks by, she stops and asks them, "So, what are you two doing for gay pride week?

Chapter 4: Quality Coding Examples

Responses that received a "3" must cite one of the two contents listed in the '3' category for each story. Examples provided for 2 and 1 are common responses provided by participants (in some cases paraphrased). Responses for "0" were fairly universal across story (incorrectly attributing a

true belief or a negative intention to the character, or otherwise mischaracterizing the intention of the statement as ironic or joking).

School Story
3: Mrs. Hayes thinks Clara is a student/a child; doesn't know Clara is a student's mother
2: Clara is short; Clara dresses plainly
1: Clara is in the school; Mrs. Hayes made a mistake

Squash Story
3: He thought his friend shared his opinions/political views; didn't know his friend was actually religious/against abortion
2: He hopes his friend will support his causes; thought his friend would follow them in their pro-choice mission
1: To describe his mission; to share his ideas; to introduce his wife

Lawyers Story
3: The secretary thought the two men were gay/ did not know they were happily married to women
2: The secretary often saw the two men together
1: The secretary is limited to her prejudices; just having fun; gossiping

Gas Story
3: He thought the old man was the attendant; didn't know the old man was another customer
2: He wanted to get his gas pumped; he was waiting and no one came to help him
1: In Pennsylvania, getting your own gas pumped was illegal; his behavior was a habit

Breakup Story
3: She doesn't know that Nicolas and Simone broke up / She thinks they're still together
2: Liana and Chloe haven't seen each other in a long time; Chloe and Jake were together [before] and Chloe had gained weight
1: Chloe gained weight; Chloe looked pregnant

Family reunion story
3: Chin Ho didn't know that his sister hurt herself/thought that she hadn't already put on makeup
2: Chin Ho wants his sister to look pretty; wants her to look her best
1: Chin Ho's sister is hurt and the makeup doesn't help

Twins story
3: She thought Jean Claude was Georges / didn't realize the person she was talking to was actually Jean Claude
2: She wants Georges to like her
1: To gossip; to express her worries

Vacation Story

3: The bellhop thought Claudia and Wayne were father and daughter; didn't know they were a married couple
2: Claudia is 25 years younger than Wayne; because of their age difference
1: To make them happy/welcome them nicely