UNIVERSITY OF CALIFORNIA SAN DIEGO

Beyond the Seen: Representation-Reality Alignment in Technology-Mediated Environments

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Cognitive Science

by

Chaolan Lin

Committee in charge:

Professor Douglas A. Nitz, Co-Chair
Professor Adena Schachner, Co-Chair
Professor Gail D. Heyman
Professor Lindsey J. Powell
Professor Federico Rossano

2026

The Dissertation of Chaolan Lin is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2026

DEDICATION

To my mother, the kindest person in the world, for her unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am deeply grateful to Doug, who supported me throughout my Ph.D. There were times when I felt uncertain, yet Doug always seemed to understand where I was, often without needing much explanation. He offered guidance that met me where I was and helped me see the next step. Looking back, his mentorship felt like a scaffold: steady, thoughtful, and always attuned to what I needed at the time. He made the Cognitive Science Department a secure harbor for me, a place of belonging and encouragement. Doug has been a role model in both scholarship and life. From conversations about doing meaningful science to discussions about the purpose of life, I learned from him in ways that extend far beyond research. For his wisdom, generosity, kindness, and reliable presence, I am deeply thankful. I would not have come this far without him. We also shared a love of dogs, a connection whose significance is difficult to put into words. I will always cherish that bond and the conversations that came with it. Wherever Tangie may be, I wish her the very best.

I am grateful to Adena Schachner in the Psychology Department for her trust and support throughout my graduate training. Her openness to new ideas helped me grow as an encouraged and independent researcher and allowed me to pursue questions my way. She created an environment in which curiosity and creativity could flourish. Beyond research, Adena provided invaluable opportunities to develop my skills, including recruiting and mentoring my own research team, navigating complex IRB processes for developmental studies, and grant writing. I have acquired extensive experience that has significantly contributed to my development as a researcher, mentor, and leader. I am grateful for the numerous ways in which she has assisted in shaping both my professional and personal growth. To my MaD Lab friends—Alexis (with whom I spent the most time and shared the most thoughts), Minju (from whom I received so

much unexpected warmth and care), Tanush, Rodney, and Shirley—and to my research assistants—Banso, Alexa, Kylie, Ray, Emma Tai, Milind, Nadeen, Vani, Olivia, Teresa, Kieran, Presley, Mary Pat, Sarah, and Emma Yang—thank you for playing an important role in my PhD research journey.

I am also grateful to Federico Rossano, Gail Heyman, Lindsey Powell, and Henrike Moll for their support, mentorship, and inspiring scholarship. Federico created opportunities for growth through the Ad Astra talk series, which became one of my favorite spaces to practice communicating research and to receive feedback from a diverse audience. Gail always offered practical advice with remarkable efficiency and consideration, and I always knew I could turn to her for guidance and support. Lindsey's active engagement in scientific discussion and genuine care for students have been truly inspiring. Her taste for spicy Chinese food is impressive. Henrike's work has been a major source of inspiration for my thinking about visual perspective-taking and social cognition more broadly. I appreciate her genuine advice and authenticity, which have significantly influenced me both professionally and personally.

I am especially grateful to Ayse Saygin, a cool and admirable neuroscientist. Her intellectual curiosity, determination, and scientific rigor continually inspire my research approach. She introduced me to ideas I have found most fascinating, including predictive coding and the cognitive mechanisms underlying the uncanny valley. As she once said, she planted a seed in my mind. That seed has grown in ways that have profoundly shaped my thinking.

I would like to thank the many incredible professors in the Cognitive Science Department who have supported and inspired my intellectual growth, especially Seana Coulson, Andrea Chiba, Marta Kutas, Haijun Xia, Gedeon Deák, David Kirsh, Amy Fox, and Bradley Voytek.

They have fostered a vibrant and inclusive community, making my graduate experience so rewarding.

I would like to thank Hadi Esmaeilzadeh, who has been a steadfast source of encouragement and affirmation, reminding me to keep moving forward and helping me recognize my strengths. I appreciate all the coaching sessions he has offered.

I would like to thank Eshin Jolly, who became an important mentor, collaborator, and role model during the final year of my Ph.D. From our very first meeting, I felt an immediate intellectual connection as our shared research interests sparked an exciting, energizing conversation. I am grateful for the time and energy he invested in our research discussions, grant proposals, and the development of future projects that helped pave the way for the next stage of my academic journey. His enthusiasm for interdisciplinary research, generosity as an advisor and collaborator, and confidence in my capabilities have left a lasting impact on me.

To my SciMinds labmates—Emma, Justin, Jinwoo (and his wife, Janice), Jacob, Aaron, Ajinkya, Christian, and Serena—thank you so much for making the lab such a joyful and supportive place to work. You are thoughtful, kind, fun, trustworthy, and inspiring as researchers, collaborators, and friends. I have cherished every shared experience with you. The memories we built together are among the most beautiful parts of my PhD journey.

To my dear graduate school friends—especially Zoe He (and Michael), Corey, Shuai, and Jane, who were my strongest sources of support during the final year of my graduate school; Huixin, Sean, Yoonwon, Lumi, Ruoying, Coxi, Xinyi, Frank, and many others—and to my cohort members—Dalin, James, Will, Vijay, Alyssa, Yifan, Andrew, Srishti, and Sam—I am deeply grateful for the friendships we built along the way.

To my family, especially my mother, whose kindness, strength, and resilience have guided me in more ways than she may ever realize. The values she lives by have had a lasting influence on the person I am today. When I was uncertain about the value of what I was doing here, she reassured me that I was doing science and contributing to human flourishing, reminding me to stay true to the purpose that brought me here in the first place. To my older sister, thank you for carrying so much of the family responsibility, allowing me the extraordinary opportunity to pursue my research in the United States. Your support made this journey possible. And to my furry baby, Alice, thank you for being my number-one fan. Your unconditional love, unwavering enthusiasm, and quiet companionship brought comfort on difficult days and joy on good ones. More than once, you gave me a reason to keep moving forward when things felt overwhelming.

There are so many people and reasons I have not been able to mention individually here. This journey would not have been possible without you, and I would not have made it this far on my own. To everyone who supported, encouraged, taught, challenged, or cared for me along the way, thank you. I wish you happiness and health in the years ahead.

Chapter 1 is currently being prepared for submission for publication: Lin, C., & Schachner, A. (In Prep). From eyes to webcam: Children learn video chat partner's visual access through self-view.

Chapter 2 is currently being prepared for submission for publication: Lin, C., & Schachner, A. (In Prep). Perceptual conflict in augmented reality promotes exploration and supports reality judgments in young users.

Chapter 3 is currently being prepared for submission for publication: Lin, C., & Schachner, A. (In Prep). From movement to meaning: perceived musical engagement in a robot enhances social evaluation of it.

VITA

2018 Master of Science in Human-Computer Interaction, Indiana University, Indianapolis

2026 Doctor of Philosophy in Cognitive Science, University of California San Diego

PUBLICATIONS

Lin, C. (2024). Parents' intention to adopt children's robots decreases after seven days of use, *16th International Conference on Social Robotics + AI (ICSR'24)*; Best short paper award

Lin, C., Šabanović, S., Dombrowski, L., Miller, A. D., Brady, E., & MacDorman, K. F. (2021). Parental Acceptance of Children's Storytelling Robots: A Projection of the Uncanny Valley of AI. *Frontiers in Robotics and AI, 8*, 49

Lin, C., Faas, T., Dombrowski, L., & Brady, E. (2017). Beyond cute: exploring user types and design opportunities of virtual reality pet games. In Proceedings of the *23rd ACM Symposium on Virtual Reality Software and Technology* (VRST)

Lin, C., Faas, T., & Brady, E. (2017). Exploring affection-oriented virtual pet game design strategies in VR attachment, motivations and expectations of users of pet games. In Proceedings of the *7th International Conference on Affective Computing and Intelligent Interaction* (ACII)

ABSTRACT OF THE DISSERTATION

Beyond the Seen: Representation-Reality Alignment in Technology-Mediated Environments

by

Chaolan Lin

Doctor of Philosophy in Cognitive Science

University of California San Diego, 2026

Professor Douglas A. Nitz, Co-Chair
Professor Adena Schachner, Co-Chair

In technology-mediated environments, perceptual information is filtered, transformed, or augmented, requiring observers to infer social, physical, and agentive states from perceptual evidence. I characterize this challenge as representation-reality alignment (RRA): The process through which humans coordinate perception, action, and inference to evaluate how perceptual

representations correspond to underlying reality and to guide behavior accordingly. Across three chapters, I examine how children and adults interpret mediated information when perceptual input provides ambiguous or indirect evidence about the world.

Chapter 1 asks how children aged 4 to 6 years infer a video chat partner's visual access from the self-view. Children increasingly used the self-view as evidence of another person's visual perspective with age; explicit reasoning emerged earlier than spontaneous action-based use. Chapter 2 asks how children aged 3 to 6 years judge whether perceptual features in augmented reality reflect physical reality or digital augmentation. Perceptual conflict increased active exploration, and exploration supported more accurate reality judgments; younger children relied more on physical inspection, whereas older children engaged more in causal testing to resolve perceptual discrepancies. Chapter 3 asks how adults infer internal capacities in a highly humanlike robot from cues of musical engagement. Holding the robot's movements constant, musical-motion synchrony and contextual cues increased perceived warmth and competence of the robot, while reducing feelings of discomfort; effects were partly explained by beliefs about the robot's ability to perceive and understand music.

Together, these findings suggest that technology-mediated environments place common inferential demands on cognition, requiring observers to evaluate how perceptual information relates to underlying social, physical, and agentive states while coordinating perception, action, and inference to guide behavior. This dissertation advances RRA as a framework for understanding how people navigate technology-mediated environments.

INTRODUCTION

In everyday perception, humans experience a simple and reliable mapping from an early age: What they see corresponds to what exists in the world. This coupling between perception and reality supports learning about objects, agents, and causal structure and guides judgments and actions. In technology-mediated environments, however, this mapping is systematically disrupted. For instance, in video-mediated interaction, what is visible to a video chat partner depends on the camera perspective rather than the direct line of sight implied by the on-screen eyes. In augmented reality (AR), digitally generated features are integrated with the physical world, creating perceptual input that may appear compelling yet only partially corresponds to physical reality. These environments, therefore, altered the relationship between perception and the world, such that what is seen is no longer a transparent guide to what is real.

This disruption introduces a second challenge: If perceptual input no longer directly corresponds to reality, how can observers determine what it signifies? In technology-mediated environments, perceptual information is filtered, transformed, or selectively generated. The same perceptual signal may support multiple interpretations, depending on how it was produced and on task demands. Consequently, perceptual input can become ambiguous about its meaning, source, and reliability.

This ambiguity creates a third problem: How should perceptual information be used to guide action? Because the same perceptual signal may support multiple mappings to underlying reality depending on context and task demands, observers may evaluate perceptual input as evidence, integrating prior knowledge and using action to determine how representations correspond to underlying social, physical, and agentive states. In this sense, perception becomes

part of an inferential process in which observers actively reason about what perceptual signals correspond to and when they can be relied upon.

I characterize this set of challenges as the problem of representation-reality alignment (RRA): The process through which humans coordinate perception, action, and inference to evaluate how perceptual representations correspond to underlying reality and to guide behavior accordingly, particularly when perceptual input is systematically transformed in technology-mediated environments. In this dissertation, I use the term *representation* to refer primarily to perceptual information, such as self-views, augmented objects, or robot behavior, that provide evidence about the possible states of the world (social, physical, or agentive). Specifically, RRA is not merely about recognizing representations or forming beliefs, but about using representations as actionable cues for inferring and responding appropriately to the social and physical world. Because the same perceptual representation can support multiple interpretations, RRA is the process of determining which interpretation best corresponds to reality.

Representations are inherently relational: They stand for something in the world while remaining distinct from the thing they represent (Deloache, 2004). Learning to reason about representations poses a well-documented challenge for young children, who often struggle to appreciate this dual status, particularly when representations closely resemble their referents or possess salient surface features (DeLoache, 1987). For example, young children may interact with pictures or scale models as though they were real objects rather than sources of information about something else (DeLoache, 2000). Children must also learn that perceptual appearances can be misleading and that objects may possess properties that diverge from their appearances (Flavell, 1986). These findings suggest that successful reasoning requires more than recognizing that something is a representation; it also requires understanding how representations relate to

underlying reality. This challenge becomes especially important in mediated environments, where perceptual input may only partially correspond to the world. Successful interaction, therefore, would depend on flexibly interpreting what perceptual signals represent, evaluating their reliability, and determining how they should guide behavior under specific task demands.

Perceptual input often provides indirect and probabilistic cues that must be interpreted in relation to context rather than treated as transparent access to reality (Clark, 2019; Friston, 2010). RRA involves learning how to use perceptual information as evidence for underlying states. For instance, when should a live self-view in a video chat be treated as an image of oneself, or as evidence of what a video chat partner can see? In AR, where perceptual information combines reality-tracking information with system-generated features, when and how does an observer reason about which aspects of the representation correspond to stable properties of the physical world and which arise from digital augmentation? In both cases, representations do not inherently specify what they denote; their meaning depends on how perceptual signals are generated, transformed, and used within a particular task context. Such inferential demands align with broader research on theory of mind and causal reasoning, in which individuals infer latent mental or physical states from incomplete, ambiguous, or indirect evidence (Gopnik & Wellman, 2012; Gweon et al., 2010; Wellman, 2014).

Theories of embodied and enactive cognition emphasize that cognitive processes are grounded in action-perception loops, in which humans generate hypotheses about the world and revise them based on the consequences of their actions (Gibson, 1979; Smith & Gasser, 2005; Varela et al., 2018). Within this framework, aligning representations with reality involves using action to probe the relationship between what is perceived and what it signifies. For instance, users may reposition an object and monitor a self-view to assess whether it becomes visible to a

video chat partner, or manipulate an object to evaluate whether changes in a screen display reflect a stable physical property or a digitally generated augmentation. Such behaviors reflect an active process of calibration in which observers iteratively test, revise, and refine the correspondence between mediated representations and underlying reality.

Research on symbolic understanding, appearance-reality distinctions, and inferential reasoning has established that children can represent, interpret, and reason about information that does not directly correspond to the world (DeLoache, 2000; Flavell et al., 1983; Gopnik & Wellman, 2012; Gweon et al., 2010; Wellman, 2014). However, these traditions have largely developed in parallel and have often examined these processes in relative isolation or in contexts where perception provides stable access to underlying reality. Technology-mediated environments introduce a different challenge: Perceptual information may be systematically transformed, indirect, or distributed across representational systems and perspectives. In such contexts, successful behavior depends not only on recognizing representations, distinguishing appearance from reality, or inferring hidden states as separate cognitive challenges, but also on coordinating these processes in real time to reason about how perceptual input maps onto underlying social or physical reality.

These considerations suggest that RRA is a central yet underexplored social-cognitive problem in technology-mediated environments. Emerging technology systems provide perceptual input that is decoupled from direct experience, transformed by representational processes, and distributed across multiple perspectives. Understanding how humans, especially children, interpret and respond to this input requires a framework that explains how perception, actions, and representational understanding coordinate in these circumstances. The following sections will outline a theoretical framework focused on these key questions.

1. Action-perception coupling in mediated environments

Embodied and enactive approaches to cognition characterize perception as an active process that emerges through ongoing interaction with the environment. These frameworks emphasize action–perception loops, in which agents generate expectations, act on the world, and revise those expectations in light of the consequences of their actions (Gibson, 1979; Smith & Gasser, 2005; Varela et al., 2018). Cognition is thus understood as embodied (grounded in the physical capacities of the body) and enacted (emerging through real-time engagement with the environment). Understanding the world depends on discovering how actions generate perceptual consequences and how perceptual information can guide future behavior.

Within this framework, perceptual information is structured by affordances, opportunities for action that emerge through the relation between the agent and the environment (Gibson, 1979). Affordances arise from the interaction between the environment and the agent's capacities, shaping which actions become possible and meaningful. Learning involves discovering stable relationships between action and perceptual feedback. Early in development, this process often unfolds through direct physical interaction. Infants and young children explore the physical environment by manipulating objects and moving through space, gradually learning how their actions transform what they perceive and what they can do (Smith & Gasser, 2005).

Technology-mediated environments change these ordinary action-perception contingencies. During joint attention over video chat, for example, successfully directing a video chat partner's attention to an object in one's own physical space depends on both where the object is positioned and how it appears through the webcam. A user who holds an object toward a partner's on-screen face would fail to establish shared attention because visibility follows the webcam's perspective rather than the line of sight implied by the partner's on-screen image. As a

result, successful action depends on understanding how actions are transformed through technological systems.

Developmental research suggests that adapting to these transformed contingencies poses challenges for young children. Children often exhibit "media errors," such as attempting to grasp or manipulate objects depicted on screens or responding to televised images as though they were physically present, suggesting difficulty coordinating actions with mediated perceptual displays (Barr, 2013; DeLoache et al., 2010; Rosengren et al., 2021; Troseth & DeLoache, 1998). Research on the video deficit effect further shows that young children experience difficulty learning from and acting on screen-based information (Troseth et al., 2006). Together, these findings suggest that children do not initially treat mediated environments as governed by the same action-perception relations as the physical world, and that they must learn how actions and perceptual consequences are reorganized by representational technologies.

A related challenge arises in AR, where perceptual feedback depends jointly on physical objects, user actions, and system-generated processes. Actions such as moving an object, changing viewpoint, or repositioning oneself can affect what becomes visible on a display in ways that differ from ordinary physical experience. Successful interaction requires learning how perceptual consequences emerge from interactions among users, physical environment, and technology systems.

These transformed contingencies may pose particular challenges for young children. Research on children's reasoning about fantastical and digitally manipulated content suggests that perceptual vividness and contextual realism can lead children to accept visually compelling yet physically implausible events as real, particularly when cues to artificiality are subtle (Harris, 2000; Troseth et al., 2019; Woolley & E Ghossainy, 2013). Children also do not consistently

prioritize physical plausibility when perceptual and contextual cues conflict (Woolley & Van Reet, 2006). In technology-mediated environments, such tendencies may make it more difficult to determine how actions relate to perceptual outcomes when those outcomes are shaped by both physical and representational processes.

Across social (e.g., video chat) and physical (e.g., AR) domains, a key developmental challenge involves learning new action-perception mappings. In video chat, successful communication depends on understanding how actions become visible through a webcam perspective. In AR, successful interaction depends on understanding how movement and changes in viewpoint affect perceptual displays. Development involves increasing flexibility in adapting to environments in which perceptual consequences are transformed, distributed across perspectives, or partly generated by technology systems.

These considerations provide a foundation for RRA. If perception emerges through interaction, then technology-mediated environments require observers to recalibrate the relationship between actions and perceptual consequences. Learning these transformed contingencies, however, is only part of the challenge. Even when observers understand how actions affect mediated displays, they must still determine what the resulting perceptual information represents. Addressing this challenge requires understanding how mediated perceptual information functions as a representation of the world.

2. Understanding representations in mediated environments

Although transformed action–perception contingencies shape how information becomes available in mediated environments, successful adaptation also depends on understanding what mediated perceptual information represents. In these environments, perceptual experiences are often generated, filtered, or transformed by technology systems. As a result, what appears on a

screen or display cannot always be interpreted in the same way as information obtained through direct perception. Users must determine how perceptual information relates to the world and what aspects of reality it depicts.

This challenge builds on a broader developmental problem concerning representational understanding. Representations stand for aspects of the world while remaining distinct from what they present (DeLoache, 2004). Young children often experience difficulty reasoning about this dual status, particularly when representations closely resemble their referents or possess salient perceptual features (DeLoache, 1987, 2000). For example, children may interact with pictures or scale models as though they were the objects themselves rather than representations of something else. Children must also learn that appearances do not always correspond directly to reality and that perceptual experiences can sometimes be misleading (Flavell, 1986). These findings suggest that successful reasoning requires understanding of not only that something is a representation, but also how that representation relates to what it depicts.

Technology-mediated environments introduce new forms of representational ambiguity. In many cases, perceptual information does not transparently specify what it refers to. The same perceptual display may support multiple interpretations depending on the context in which it is encountered and the purpose for which it is used. Interpreting mediated information requires determining what a representation denotes and how it should be understood within a particular situation.

Video-mediated interaction illustrates this challenge. A self-view is, at one level, an image of oneself. At the same time, it can be interpreted as information about how one's actions are displayed through the communication system. The same representation can thus serve

different functions depending on the demands in a given context and what aspect of the interaction it represents.

AR introduces a different representational challenge. Digitally generated and reality-tracking features may appear seamlessly integrated within the same perceptual scene, making it difficult to determine how individual features correspond to the physical environment. Rather than evaluating an entire scene as either real or artificial, users may need to interpret the origins of particular perceptual features and determine how those features relate to the world beyond the display. The challenge lies in understanding how different components of a single perceptual experience correspond to different sources.

Across both video chat and AR, successful interaction depends on understanding the relationship between perceptual representations and what they depict. Users must determine what mediated information refers to, how it relates to the world, and which interpretation is appropriate for the current context. This challenge extends beyond recognizing that something is a representation. It involves flexibly interpreting representations whose meanings may shift across situations and uses.

Yet identifying what a representation depicts does not fully resolve uncertainty. A representation may still be compatible with multiple possible states of the world. Observers must decide which interpretation best explains the current situation and how the available information should be used. This introduces a further challenge: using mediated perceptual information to reason about aspects of reality that cannot be directly observed.

3. Inferring hidden states from perceptual evidence

Once perceptual information has been interpreted as representational, a further challenge emerges: determining what aspects of reality it reveals. In mediated environments, perceptual

signals often provide incomplete, indirect, or ambiguous information about the world. The same representation may be consistent with multiple possible states of affairs, requiring observers to move beyond what is immediately perceptible and infer the underlying causes of what they perceive.

This challenge aligns with broader theories of cognition that characterize perception and reasoning as forms of inference under uncertainty. Rather than passively receiving information from the environment, observers actively interpret perceptual input by generating and evaluating hypotheses about its causes (Clark, 2019; Friston, 2010). Because perceptual signals are often underdetermined, successful interpretation requires integrating perceptual information with contextual knowledge and prior expectations to identify the most plausible explanation for what is observed (Xu & Kushnir, 2013).

Humans begin to engage in this form of reasoning early in development. Even infants use patterns of evidence to infer hidden properties of objects and agents, including causal relations, physical support structures, and goal-directed behavior (Baillargeon, 2004; Spelke, 2022). In social cognition, children infer mental states such as beliefs, intentions, and perceptual access from observable behavior (Wellman, 2014). These findings suggest that humans are capable of reasoning beyond immediately available information and using perceptual input to infer aspects of the world that cannot be directly observed.

Technology-mediated environments place additional demands on this inferential process because the relationship between perceptual information and reality is often indirect. Observers must determine which aspects of a representation are informative, evaluate alternative interpretations, and update their conclusions as new information becomes available through interaction. RRA concerns how these inferential processes operate when relevant social,

physical, or agentive states are not directly accessible and must be reconstructed from mediated perceptual information.

Video-mediated interaction provides one example. A communication partner's visual access cannot be directly observed because it depends on how actions are captured and transmitted through a webcam and interfaces. Users must infer what another person can see from information provided by the mediated system and adjust their behavior accordingly. Success depends not only on recognizing the representational functions of the display but also on using that information to infer another person's current perceptual state.

AR presents a related inferential challenge in the physical domain. Because digitally generated and reality-tracking features are integrated within the same perceptual scene, observers must determine which aspects of their experience correspond to stable properties of the physical environment and which arise from digital augmentation. Resolving this ambiguity often requires comparing alternative interpretations and evaluating how perceptual information changes across viewpoints, contexts, and interactions.

Although these examples differ in content, they share a common inferential structure. Perceptual information provides only partial access to the relevant state of the world. In video chat, the target of inference is another person's perceptual access. In AR, the target of inference is the reality status of perceptual features. In both cases, observers must move beyond what is directly perceived and determine which interpretation best accounts for the available information.

Actions play an important role in this process because they generate new information that can help distinguish among competing interpretations. By manipulating objects, changing viewpoints, or modifying their behavior, observers can observe how mediated representations

respond and use the resulting information to refine their understanding of the situation. In this sense, inference is not merely a cognitive process but an interactive one, unfolding through ongoing coordination between perception and action.

This perspective suggested that RRA is fundamentally an inferential challenge. Successful adaptation to mediated environments requires more than learning new action-perception contingencies or recognizing representational relationships. It requires using mediated perceptual information to infer underlying social, physical, and agentive states under conditions of uncertainty. Developmental change may therefore reflect increasing flexibility in how children evaluate perceptual information, generate interpretations, and revise those interpretations in response to new evidence.

4. Developmental emergence of RRA

This dissertation focuses on early childhood, specifically ages 3 to 6 years, because this period marks a developmental transition during which many of the component processes relevant to RRA are already present but undergoing substantial reorganization. Between these ages, children show rapid improvements in perspective-taking, representational understanding, appearance-reality reasoning, causal inference, and action-based learning (DeLoache, 2000; DeLoache, 2004; Gopnik & Wellman, 2012; Moll & Tomasello, 2006; Wellman, 2014). Children become increasingly capable of reasoning about hidden states, interpreting representations flexibly, and using action to test hypotheses about the world (Schulz, 2012; Xu & Kushnir, 2013). At the same time, these abilities remain context-dependent and are not yet fully integrated, making early childhood a theoretically informative period for examining how the capacities underlying RRA become coordinated.

Although developmental research has extensively examined symbolic understanding, perspective-taking, causal reasoning, and embodied learning, existing accounts provide limited insight into how children coordinate these abilities in technology-mediated environments. Several developmental possibilities remain. One possibility is that developmental change reflects increasing knowledge of representational systems, allowing children to reason more effectively about how webcams, displays, or AR interfaces transform perceptual information. A second possibility is that developmental change reflects increasing flexibility in interpreting perceptual input, enabling children to treat mediated representations as informative but potentially imperfect indicators of reality. A third possibility is that developmental change reflects increasing use of action to evaluate competing interpretations, with children becoming more likely to explore, test, and revise their understanding of how perceptual information relates to the world. Although these possibilities are not mutually exclusive, they make distinct predictions regarding how children would reason and behave in mediated contexts.

I propose that developmental change in RRA reflects increasing flexibility in how children interpret and use perceptual information under uncertainty. Specifically, younger children are more likely to treat perceptual input as a relatively direct reflection of reality, relying on immediate appearances to guide judgment and action. Older children, in contrast, increasingly evaluate perceptual information in relation to context, consider alternative interpretations, and use perceptual information to reason about underlying social and physical states. This developmental shift may emerge without explicit knowledge of how technological systems operate, such as the mechanics of webcams or AR interfaces. Instead, developmental change reflects increasing ability to coordinate action, representation, and inference when perceptual information provides only indirect access to reality.

This account generates several developmental predictions. Younger children should rely more heavily on immediate perceptual appearances, whereas older children should become increasingly capable of treating perceptual information as evidence that can be interpreted in multiple ways. Age-related improvements should be particularly evident when perceptual cues are transformed, incomplete, or conflicting, conditions that require evaluating alternative interpretations rather than relying on direct perception alone. Older children should also become more likely to use exploratory actions strategically, generating new information that helps distinguish among competing explanations and reduce uncertainty.

The present dissertation evaluates these predictions across multiple technology-mediated contexts. Chapter 1 examines how children use self-view information during video chat to reason about a partner's visual access. Chapter 2 investigates how children evaluate self-view information to reason about the reality status of perceptual features in AR and how exploration contributes to those judgments. Together, these studies examine whether developmental change reflects increasing flexibility in how perceptual information is interpreted, evaluated, and used to guide behavior when relevant aspects of reality are not directly observable.

More broadly, this framework suggests that developmental change in RRA may not primarily involve acquiring knowledge about technologies but reflect a more general shift in how children relate perceptual information to reality. As children develop, they increasingly move beyond treating perception as a contextualized reflection of the world and become more capable of using perceptual information as evidence for reasoning about underlying social and physical states.

5. Extending RRA beyond childhood

The challenges described thus far are not unique to childhood. As new technology becomes increasingly embedded in everyday life, people of all ages encounter situations in which perceptual information offers only indirect access to the underlying reality. Modern technology frequently generates perceptual experiences that require interpretation, whether through screens, augmented displays, virtual environments, or artificial agents. In each case, users must determine how perceptual information relates to the world and what conclusions can reasonably be drawn from it.

This challenge extends naturally to judgments about artificial agents. Unlike physical objects or human interaction partners, artificial agents often exhibit behaviors that appear socially meaningful yet provide limited information about the capacities that generate them. Observable behavior may suggest perception, understanding, intentionality, or engagement, yet these internal capacities cannot be directly observed. As a result, people must infer hidden agentive states from perceptual evidence in much the same way that they infer hidden social or physical states in other mediated environments.

Artificial agents provide a useful extension of the RRA framework because they introduce uncertainty not about what is physically present or what another person can see, but about the nature of the agent itself. A conversational response may reflect genuine understanding, statistical pattern matching, or a scripted sequence. Coordinated movement may indicate meaningful engagement, a programmed response, or a purely mechanical process. Observable behavior does not uniquely specify which interpretation is correct. Users must determine how behavioral cues relate to the underlying capacities of the system.

This challenge shares the same inferential structure described in earlier sections. In video chat, observers infer a partner's visual access from mediated information. In AR, observers infer

whether perceptual features correspond to physical reality or digital augmentation. In interactions with artificial agents, observers infer internal capacities such as perception, understanding, intentionality, and social engagement. Across all three domains, perceptual information provides only partial evidence about states that are not directly observable.

The agentive domain offers an opportunity to examine whether the principles underlying RRA generalize beyond childhood and beyond questions of social and physical reality. If RRA reflects a broader capacity to interpret perceptual information, evaluate competing interpretations, and infer underlying states under conditions of uncertainty, then similar processes should shape how adults understand artificial agents. Observable behavior becomes perceptual evidence, interpretations of behavior become representational hypotheses, and judgments about internal capacities become inferential conclusions.

The present dissertation begins this line of extension by examining adults' evaluations of a highly humanlike robot. Although young children can reason about mental states, studying agentive RRA in childhood presents substantial developmental and methodological challenges. Young children often attribute animacy, intention, and mental capacities broadly, particularly when movement appears contingent or socially meaningful (Gelman & Opfer, 2007; Jipson & Gelman, 2007; Johnson et al., 1998; Severson & Carlson, 2010). Adults provide a more theoretically tractable starting point because they possess more stable knowledge about technology while still relying on perceptual information to infer hidden capacities.

Chapter 3 examines whether cues of musical engagement influence how adults interpret a robot's behavior and evaluate it socially. More broadly, the chapter asks whether judgments of artificial agents depend not only on what observers perceive, but also how they interpret the relationship between observable behavior and underlying capacities. By extending RRA to the

agentive domain, this dissertation investigates whether a common inferential framework can explain how people reason about hidden, social, physical, and agentive states across a diverse range of technology-mediated environments.

6. Dissertation outline

This dissertation investigates how humans use perceptual information to infer underlying social, physical, and agentive states in technology-mediated environments. Guided by the RRA framework, I examine how observers interpret representations, evaluate perceptual evidence, and coordinate action when relevant aspects of reality are not directly observable. Across three chapters, I investigate complementary forms of RRA across childhood and adulthood.

Chapters 1 and 2 focus on the developmental emergence of RRA during early childhood, a period in which children increasingly move beyond treating perception as a direct reflection of reality and become more capable of using perceptual information as evidence about hidden states. These chapters examine complementary developmental challenges in the social and physical domains, highlighting how technology-mediated environments transform the relationship between perception and reality.

In Chapter 1, I examine RRA in the social domain, asking how children aged 4 to 6 years infer a video chat partner's visual access from self-view during remote interaction. In video chat, another person's perspective is no longer specified through shared physical space but is filtered through a camera and display. Children must therefore determine how perceptual representations correspond to another person's perceptual access and use this information to guide their actions. Specifically, this chapter investigates whether children use the self-view as a cue for reasoning about a video chat partner's visual access and how this ability changes with age. Findings indicate that children increasingly use self-view information to infer another person's visual

access across development. When self-view was available, children were more likely to look at it, and position objects relative to the webcam, suggesting that they treated self-view as informative about a partner's perspective. Scaffolding further supported performance, revealing developmental progression in how children coordinate mediated perceptual cues with action. Explicit reasoning revealed that children selectively treated self-view as relevant for visual access but not for action-based judgments, indicating an early understanding of the informational significance of self-view.

In Chapter 2, I examine RRA in the physical domain, asking how children aged 3 to 6 years judge whether perceptual features in AR correspond to physical reality or digital augmentation. Unlike video-mediated interaction, where the challenge lies in inferring another person's perceptual state, AR introduces ambiguity regarding the reality status of perceptual features themselves. Children must evaluate conflicting information and determine which aspects of their perceptual experience correspond to the stable properties of the physical world. Specifically, this chapter investigates whether perceptual conflict drives active exploration and how exploratory behavior contributes to reality judgments in AR. Findings show that, when faced with conflicting visual information across views, children generate and evaluate evidence by acting on their environment: they move objects, shift viewpoints, and monitor resulting changes in mediated displays. Perceptual conflict increased exploratory behavior, and exploration predicted more accurate reality judgments. Children also showed increased exploration following exposure to conflict, suggesting a short-term recalibration in how mediated perceptual information was evaluated.

In Chapter 3, I extend RRA beyond childhood and into the agentive domain, examining how adults infer internal capacities from observable robot behavior. Humanlike robots often

exhibit behaviors that appear socially meaningful while providing only indirect information about the capacities that generate them. Observers must therefore determine how behavioral cues relate to underlying agentive states such as perception, understanding, intentionality, and engagement. Specifically, this chapter examines whether and how cues of musical engagement influence adults' interpretations of robot behavior and their social evaluations. Findings indicate that musical-motion synchrony and contextual framing shaped perceptions of the robot's warmth, competence, and discomfort. These effects were partly explained by beliefs about the robot's capacity to perceive and understand music, suggesting that social evaluation depended on how behavioral cues were interpreted as evidence about underlying capacities.

Together, these chapters investigate a common inferential challenge across distinct technology-mediated environments: determining how mediated perceptual information relates to aspects of reality that are not directly observable. Across social, physical, and agentive domains, observers must interpret representations, evaluate perceptual evidence, and coordinate action to reduce uncertainty about the world. By examining these processes across development and adulthood, this dissertation advances RRA as a framework for understanding how humans align perception with reality in increasingly mediated environments.

REFERENCES

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, *13*(3), 89–94.

Barr, R. (2013). Memory constraints on infant learning from picture books, television, and touchscreens. *Child Development Perspectives*, *7*(4), 205–210.

Clark, A. (2019). *Surfing uncertainty*. Oxford University Press.

De Graaf, & Malle. (2017). How people explain action (and autonomous intelligent systems should too). *2017 AAAI Fall Symposium Series*. https://www.aaai.org/ocs/index.php/FSS/FSS17/paper/viewPaper/16009

De Graaf, M. M. A., & Malle, B. F. (2019). People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 239–248.

DeLoache, J. S. (1987). Rapid change in the symbolic functioning of very young children. *Science*, *238*(4833), 1556–1557.

DeLoache, J. S. (2000). Dual representation and young children's use of scale models. *Child Development*, *71*(2), 329–338.

Deloache, J. S. (2004). Becoming symbol-minded. *Trends in Cognitive Sciences*, *8*(2), 66–70.

DeLoache, J. S., Chiong, C., Sherman, K., Islam, N., Vanderborght, M., Troseth, G. L., Strouse, G. A., & O'Doherty, K. (2010). Do babies learn from baby media? *Psychological Science*, *21*(11), 1570–1574.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886.

Flavell, J. H. (1986). The development of children's knowledge about the appearance–reality distinction. *The American Psychologist*, *41*(4), 418–425.
Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology*, *17*(1), 99–103.

Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Development of the appearance--reality distinction. *Cognitive Psychology*, *15*(1), 95–120.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3), 143–166.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138.

Gelman, S. A., & Opfer, J. E. (2007). Development of the animate–inanimate distinction. In *Blackwell Handbook of Childhood Cognitive Development* (pp. 151–166). Blackwell Publishers Ltd.

Gibson, J. J. (1979). *Ecological approach to visual perception*. Houghton Mifflin.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(20), 9066–9071.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.

Harris, P. L. (2000). *The work of the imagination*. Blackwell.

Jipson, J. L., & Gelman, S. A. (2007). Robots and rodents: children's inferences about living and nonliving kinds. *Child Development*, *78*(6), 1675–1688.

Johnson, S., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. In *Developmental Science* (Vol. 1, Issue 2, pp. 233–238). https://doi.org/10.1111/1467-7687.00036

Moll, H., & Meltzoff, A. N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development*, *82*(2), 661–673.

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *The British Journal of Developmental Psychology*, *24*(3), 603–613.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine / IEEE Robotics & Automation Society*, *19*(2), 98–100.

Rosengren, K. S., Kirkorian, H., Choi, K., Jiang, M. J., Raimer, C., Tolkin, E., & Sartin-Tarm, A. (2021). Attempting to break the fourth wall: Young children's action errors with screen media. *Human Behavior and Emerging Technologies*, *hbe2.273*. https://doi.org/10.1002/hbe2.273

Schulz, L. (2012). The origins of inquiry: inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*(7), 382–389.

Severson, R. L., & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks: The Official Journal of the International Neural Network Society*, *23*(8-9), 1099–1103.

Smith, L., & Gasser, M. (2005). The development of embodied cognition: six lessons from babies. *Artificial Life*, *11*(1-2), 13–29.

Spelke, E. S. (2022). Core knowledge. In *What Babies Know* (pp. 190–200). Oxford University PressNew York.

Troseth, G. L., & DeLoache, J. S. (1998). The medium can obscure the message: young children's understanding of video. *Child Development*, *69*(4), 950–965.

Troseth, G. L., Flores, I., & Stuckelman, Z. D. (2019). When Representation Becomes Reality: Interactive Digital Media and Symbolic Development. *Advances in Child Development and Behavior*, *56*, 65–108.

Troseth, G. L., Saylor, M. M., & Archer, A. H. (2006). Young children's use of video as a source of socially relevant information. *Child Development*, *77*(3), 786–799.

Varela, F. J., Thompson, E., & Rosch, E. (2018). *The embodied mind*. MIT Press.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117.

Wellman, H. M. (2014). *Making Minds: How Theory of Mind Develops*. Oxford University Press.

Woolley, J. D., & E Ghossainy, M. (2013). Revisiting the fantasy–reality distinction: Children as naïve skeptics. *Child Development*, *84*(5), 1496–1510.

Woolley, J. D., & Van Reet, J. (2006). Effects of context on judgments concerning the reality status of novel entities. *Child Development*, *77*(6), 1778–1793.

Xu, F., & Kushnir, T. (2013). Infants are rational constructivist learners. *Current Directions in Psychological Science*, *22*(1), 28–32.

CHAPTER 1 FROM EYES TO WEBCAM: CHILDREN LEARN VIDEO CHAT PARTNER'S VISUAL ACCESS THROUGH SELF-VIEW

Video chat has become a widespread context for young children's social interaction and learning. Even in early childhood, children engage with social others through live video platforms, often showing greater engagement and learning from contingent video interactions than from pre-recorded content (McClure et al., 2018; Roseberry et al., 2014; Troseth et al., 2006). These findings suggest that foundational processes of social cognition are increasingly unfolding in environments where interaction is mediated by webcams and screens rather than shared physical space.

This shift introduces a fundamental challenge for understanding others' visual perspectives. In face-to-face interaction, children can rely on a stable mapping between perceptual cues and visual access: people see what is in front of their eyes. During video chat, however, this mapping breaks down: the partner's eyes appear on the screen, but their visual access is determined by the webcam's location and orientation. As a result, children must inhibit the well-established assumption in the physical world that vision originates from eyes and instead infer that visual access depends on an unseen device. This creates a novel inferential problem: how to identify another's visual perspective over video chat when visual access is mediated by webcams.

On the other hand, unlike face-to-face interaction, where others' perspectives must be inferred indirectly (Moll & Meltzoff, 2011a), video chat provides a potential means of resolving this problem. Modern video chat platforms often offer a self-view window that displays the scene as it is captured by the webcam, providing a concrete, continuously updated representation

of what another person can see. We propose that, once understood, this representation can serve as a source of feedback, allowing children to infer their partner's visual perspective.

Specifically, we hypothesize that children can use the self-view to learn that visual perspective does not depend on on-screen eyes. If children recognize that their self-view reflects what their partner can see, they can use this to guide their behavior in real time. For instance, to ensure their partner has visual access to an object from their side, they can refer to the self-view and confirm that an object is shown there. In this way, the self-view may enable children to replace a default, gaze-based strategy with a non-eyes-based understanding of visual access, supporting successful coordination (e.g., joint attention) in video chat.

However, this learning process is constrained by two core aspects of cognitive development: visual perspective-taking and symbolic understanding. First, children must be able to reason about what others can see in the physical world. Although even young children understand that visual access is linked to perception, more robust and flexible perspective-taking emerges gradually, with consistent success on more complex tasks typically observed around ages 4.5 to 5 (Flavell et al., 1981; Moll & Meltzoff, 2011b; Wellman, 2014). Second, children must understand the representational nature of on-screen information. The self-view is not the partner's perspective itself, but a visual representation of it; the partner's on-screen face represents the partner, not the partner themselves. Prior work shows that younger children often struggle with such dual representations, sometimes treating images as if they were the objects or people they depict (DeLoache, 2007). In the context of video communication, this limitation is evident in "media errors," such as attempting to hand objects through the screen (Rosengren et al., 2021), suggesting difficulty distinguishing between representation and reality. These developmental changes suggest that children around age four may begin to show selective

success in using self-view to reason about another person's visual access, while more consistent performance may emerge later in the preschool years.

These developmental constraints are compounded by a powerful perceptual bias: Attentional bias towards eyes and gaze direction. Sensitivity to gaze emerges early in infancy and operates rapidly and automatically (Brooks & Meltzoff, 2005; Langton & Bruce, 1999). Infants and young children readily follow gaze cues, including those presented on screens (Capparini et al., 2023), indicating that the visual system strongly prioritizes eyes as indicators of perception. In video chat, this bias may mislead children into attributing visual access to the on-screen eyes rather than to alternative sources of perception (e.g., the webcam). This predicts that younger preschoolers may continue to rely on gaze-based strategies, particularly in spontaneous situations, whereas older children may increasingly use self-view to coordinate action and reasoning about what another person can see.

These considerations raise a key theoretical question: Is children's understanding of visual perspective relatively fixed or adaptable? Modularity accounts propose that perspective-taking relies on specialized, early-emerging mechanisms that prioritize cues such as gaze direction and line of sight for inferring others' visual access (Baron-Cohen, 1997; Leslie, 1987, 1994), which would predict difficulty adapting to contexts where these cues are misleading. In contrast, constructivist accounts emphasize that children actively build and revise their understanding of others' minds through experience (Gopnik & Wellman, 2012; Wellman, 2014). From this perspective, video chat presents an opportunity for learning, in which children can use feedback from the self-view to revise expectations about how perception works in a novel environment.

We therefore predicted that children in the preschool years, particularly between 4 and 6 years of age, would begin to use the self-view as a source of information about others' visual perspectives. At this stage, children have sufficiently developed both perspective-taking and symbolic understanding to interpret the self-view to infer the partner's visual access. As a result, they should increasingly rely on the self-view to understand what the partner can see, overcoming misleading gaze cues from the on-screen eyes, and achieving successful coordination in video chat. By examining how children navigate this technology-mediated environment, the present study tested whether perspective-taking is constrained by fixed perceptual mechanisms or shaped by flexible, experience-driven learning.

In three preregistered experiments, we tested whether children aged 4.0 to 5.9 years use the self-view as a source of information about others' visual perspectives during video chat. All studies were conducted via Zoom, with children participating from home in a naturalistic setting. Prior to participation, parents completed a questionnaire about their child's prior experience with video chat. With parental assistance, the Zoom interface was standardized across participants. In the self-view condition, the child's self-view and partner-face were displayed side by side (left-right position counterbalanced). In the no self-view condition, the self-view window was hidden, and the screen only showed the partner's face.

Experiment 1 established whether children spontaneously use self-view to infer what a video chat partner can see. Children completed a show-and-tell task in which they were asked to show small toys to a partner over video chat (4 trials; one toy per trial). We manipulated whether the self-view was available for children (self-view turned on vs. off, between-subjects).

We predicted that children who interpret the self-view as informative of the partner's visual perspective would use it to guide their actions. Specifically, when the self-view is

available, these children were expected to position objects in front of the webcam so that the object appears in view. In contrast, children relying on gaze-based cues were expected to present objects toward the partner's on-screen eyes, particularly when the self-view was unavailable (see Figure 1.1).

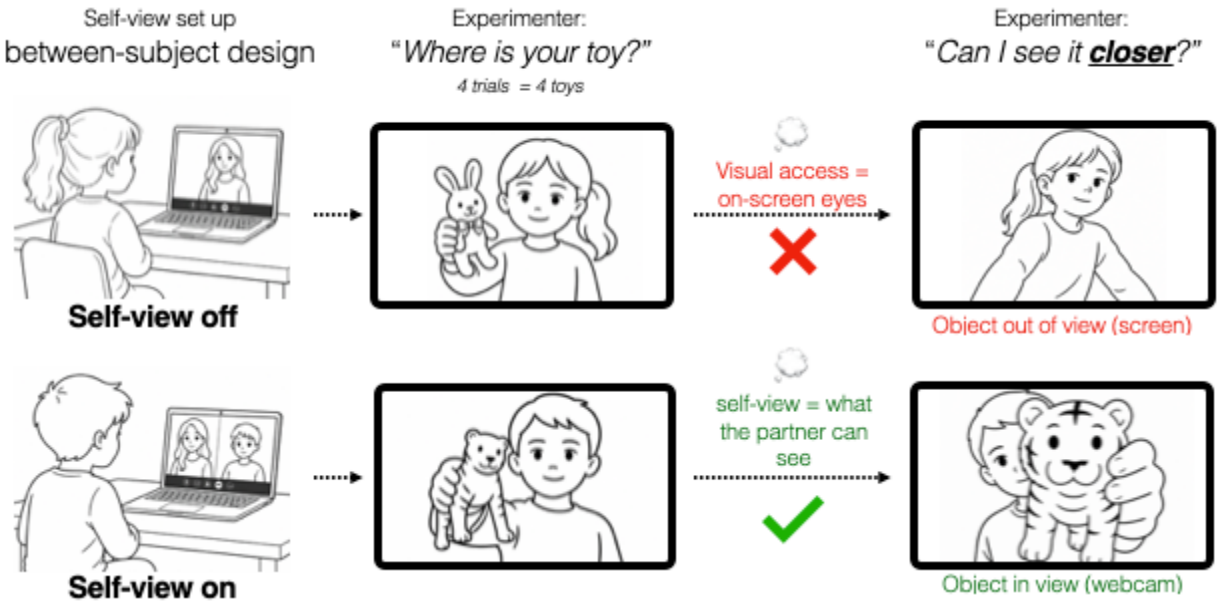

Figure 1.1: Experimental paradigm of Experiment 1. Children completed a show-and-tell task during video chat with self-view either available or unavailable (between-subjects). The task tested whether children used self-view to infer what a partner could see, reflected in whether objects were positioned toward the partner's on-screen face or in front of the webcam.

To investigate whether children can adapt their understanding of visual access in video chat, **Experiment 2** tested whether children could use self-view as representational feedback to revise how they infer their partner's visual access when provided with structured scaffolding. The self-view was available for all children. Each child completed a task in which they were asked to block their partner's visual access during video chat. Critically, successful performance required identifying the webcam as the source of visual input. To assess learning, we used a hierarchical scaffolding paradigm with three progressively more explicit levels of support,

progressing from spontaneous performance to emulation and imitation, forms of observational learning that differ in whether children observe outcomes or actions (Horner & Whiten, 2005; Tomasello et al., 1993). Children advanced to the next level only if they failed at the preceding stage. The task began with a spontaneous condition in which children were asked to block the partner's view ("Can you block me from seeing you?"), providing the goal without additional support. If unsuccessful, children proceeded to an emulation condition in which children covered their eyes while the experimenter covered the webcam, allowing them to observe the outcome of blocked visual access without seeing the action used to produce it. If children continued to fail, they proceeded to an imitation condition in which they observed the experimenter's hand approach the screen-mounted webcam area and cover the webcam. This design allowed us to test whether children could shift from a default gaze-based strategy toward greater use of webcam-based cues as representational support increased.

We predicted that children who recognize self-view as informative of their partner's visual access would be more likely to block the webcam, particularly when scaffolding provided clearer outcome and action information. In contrast, children relying on gaze-based cues were expected to continue covering the partner's on-screen face, with children who did not succeed spontaneously showing greater success as support increased. (see Figure 1.2).

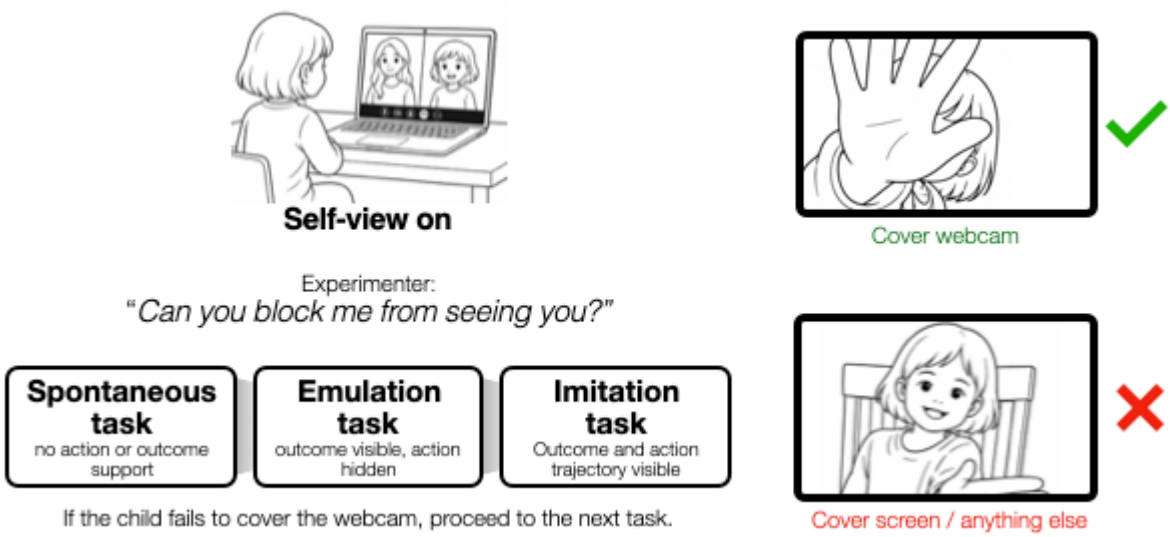

Figure 1.2: Experimental paradigm of Experiment 2. Children completed a video-chat task in which they were asked to block a partner's visual access while self-view was available. Successful performance required covering the webcam. Children progressed through a hierarchical scaffolding paradigm with increasing levels of support: spontaneous performance (goal provided without support), emulation (outcome visible, action hidden), and imitation (action trajectory visible). The paradigm tested whether children could adapt their behavior by using self-view to identify the webcam as the source of visual input.

To test whether this understanding extends beyond action-based coordination to explicit reasoning, **Experiment 3** examined whether children could reason about a video chat partner's visual access in a more abstract verbal task. The self-view was available for all children. Children and the experimenter each wore different shapes on their foreheads, such that neither could directly see their own shape and successful performance depended on reasoning from screen-based information. Children were asked questions about what the experimenter could see, touch, or see but not touch. The see question assessed whether children could infer what was visually available to the partner through video chat, whereas the touch question served as a control condition grounded in physical access. The see-but-not-touch question tested whether children distinguished the partner's visual perspective from all information visible on the screen, addressing the possibility that some children might assume the partner also had access to their

own self-view. This design tested whether children could use self-view to reason explicitly about another person's visual perspective beyond immediate action coordination.

We predicted that children who recognize self-view as informative of their partner's visual perspective would accurately answer the see and see-but-not-touch questions by distinguishing the partner's visual access from all information visible on the screen (self-view and partner view). In contrast, children who do not yet interpret self-view as informative of the partner's perspective were expected to perform less accurately on visual-access questions and rely more on alternative cues, such as the partner's on-screen face or the full set of visible information (see Figure 1.3).

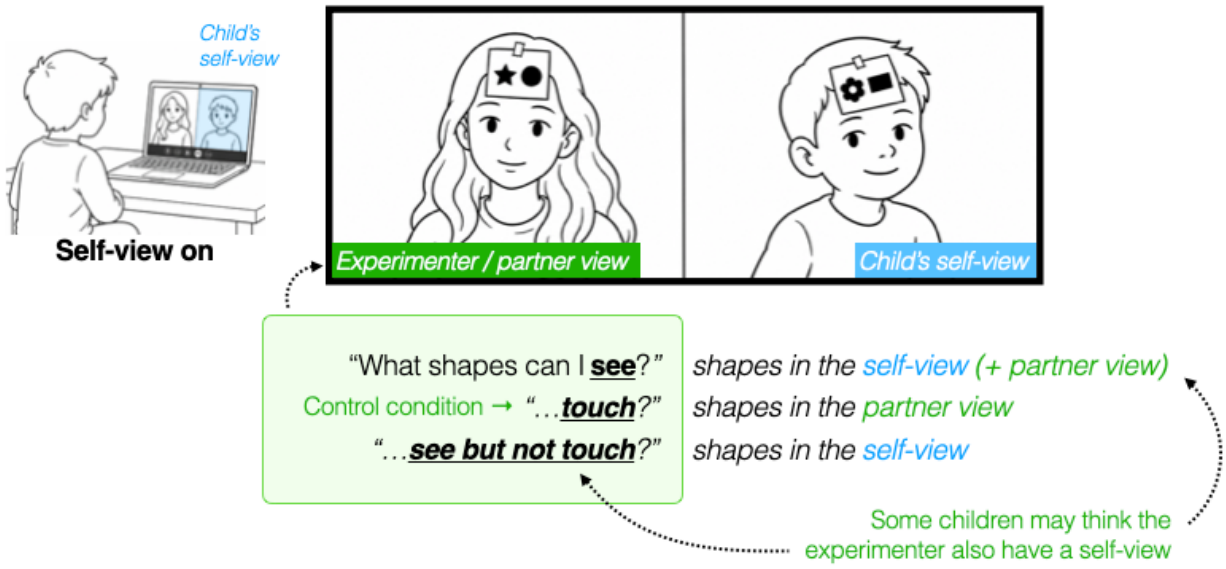

Figure 1.3: Experimental paradigm of Experiment 3. During a video-chat reasoning task, children and the experimenter wore different shapes on their foreheads while a self-view was available. Children completed naming and counting tasks involving questions about what the partner could see, touch (control condition), or see but not touch. The task tested whether children could use self-view to infer a partner's visual access and distinguish it from information merely visible on the screen.

Results

Experiment 1.

To test whether access to the self-view supports children's understanding of their video chat partner's visual perspective, we compared reaching behavior when showing objects to the partner across conditions (self-view on vs. off). Children showed a clear shift in behavior depending on whether self-view was available. In the self-view-off condition, children almost exclusively attempted to show objects by holding them up to the on-screen image of the partner (M= 94.6% of trials per child, SE = 2.45%, 95% CI [89%–0.99%]). This resulted in a failed attempt to initiate joint attention because the object remained outside the webcam's field of view. In contrast, when self-view was available, children were substantially more likely to correctly reach toward the webcam (M = 53.8%) than in the self-view-off condition (M = 5.4%), Welch's $t(91.9) = 7.78$, $p < .001$. Confidence intervals were estimated using participant-level bootstrap resampling (10,000 samples, percentile method).

We next tested whether the effect of the self-view on webcam-reaching behavior varied with age using mixed-effects logistic regression with random intercepts for participants. The model revealed a strong main effect of condition: children were substantially more likely to reach toward the webcam when self-view was available, $b = 11.59$, $SE = 2.52$, $z = 4.60$, $p < .001$. This effect was moderated by age, as indicated by a significant condition × age interaction, $b = 14.74$, $SE = 4.64$, $z = 3.18$, $p = .001$. Model predictions showed that, in the self-view-on condition, the probability of webcam reaching increased with age, reaching 50% at approximately 4.95 years and approaching 92% at the upper end of the age range. In the self-view-off condition, webcam reaching remained consistently low across ages, with predicted probabilities not exceeding 10%. These findings suggest that age-related change was specific to

contexts in which the self-view was available. When self-view was absent, even the oldest children rarely reached toward the webcam, indicating limited use of webcam location without self-view support. By contrast, when self-view was available, older children increasingly positioned objects in front of the webcam, consistent with greater use of self-view to guide object placement during video chat. See Figure 1.4.

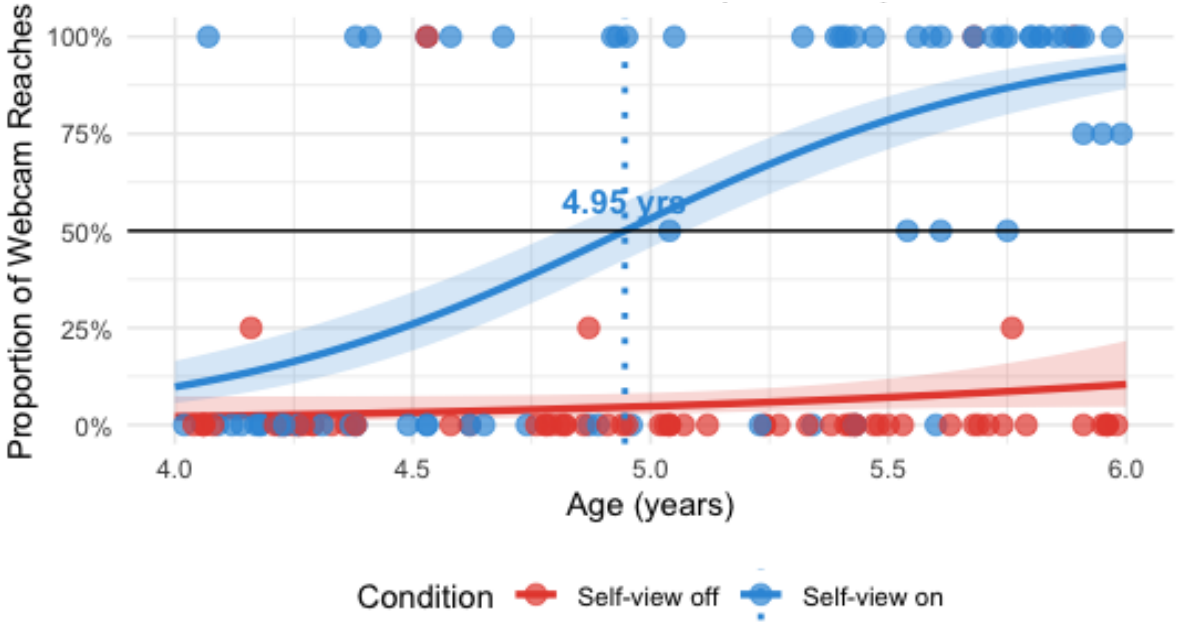

Figure 1.4: Developmental change in the use of the self-view to guide social interactions. When the self-view was absent, even 6-year-old children failed to show objects by reaching for the webcam (red). When the self-view was present, older but not younger, children showed greatly improved performance, successfully showing an object by reaching toward the webcam (blue). From ages 4-6, children increasingly use their self-view as a source of information about their partner's visual perspective.

Looking proportions to the self-view (LP-SV) and to the partner's face (LP-PF) were computed as the proportion of time spent fixating each region relative to total looking time within a trial (excluding experimenter prompts). A gaze preference index (LP-C) was defined as the difference between LP-SV and LP-PF. At the participant level, gaze preference was strongly

associated with webcam-reaching behavior, $r = .68$, 95% CI [.53, .79], $p < .001$. However, this association was largely accounted for by age. In a regression model including both predictors, age remained a significant predictor of webcam-reaching behavior, $\beta = .25$, $p = .004$, whereas gaze preference was not, $\beta = -.002$, $p = .77$. This pattern indicates the association between age preference and webcam-reaching behavior primarily reflected developmental differences. See Figure 1.5.

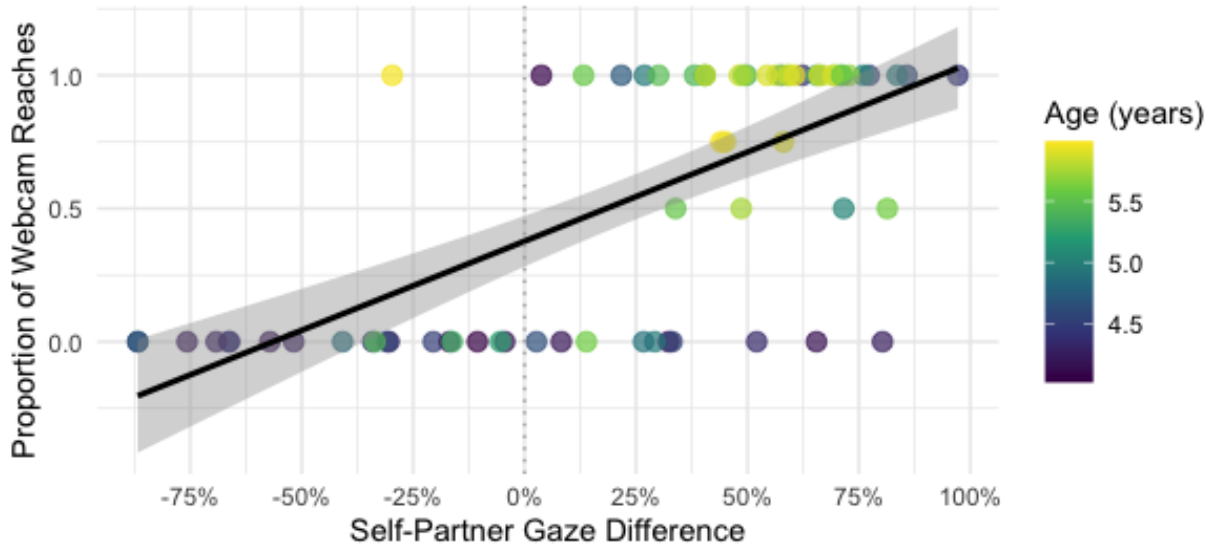

Figure 1.5: Association between gaze and webcam reaches in the self-view condition. The x-axis shows the gaze preference index (LP-C = LP-SV − LP-PF), with positive values indicating greater attention to the self-view relative to the partner's face. The y-axis shows the proportion of webcam reaches. Each point represents a participant and is color-coded by age. Greater self-view preference was associated with higher rates of webcam reaching ($r = .68$, $p < .001$), although this association was largely accounted for by age-related developmental differences.

Experiment 2.

To examine children's initial use of device-relevant strategies, we modeled the first device-relevant response (partner face, self-view, or webcam) using a multinomial logistic regression, with age (mean-centered) and scaffolding stage as predictors. Age significantly

predicted children's initial strategy selection, likelihood-ratio $\chi^2(2) = 16.50$, $p < .001$. Relative to partner-face–directed responses, older children were more likely to initiate webcam-directed responding, $b = 1.80$, $SE = 0.75$) and less likely to engage with the self-view, $b = -1.44$, $SE = 0.98$. See Figure 1.6.

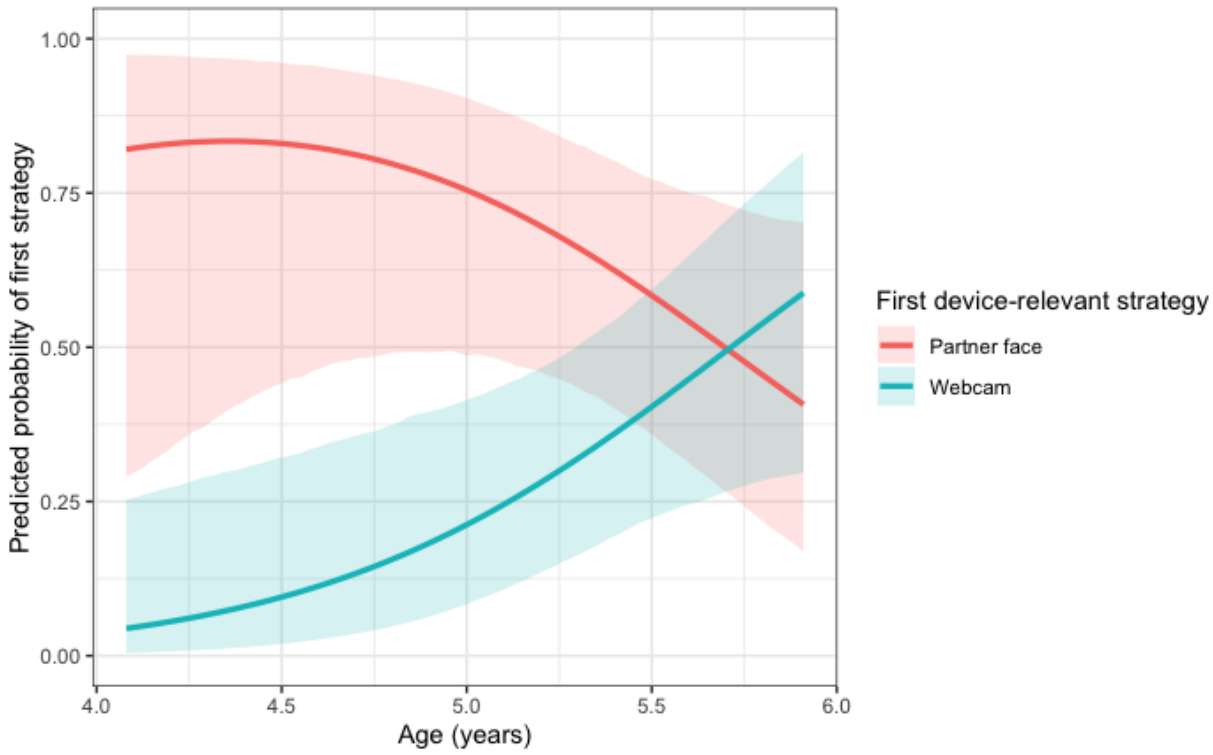

Figure 1.6: Model-predicted probability of children's reaches (partner-face vs webcam) by age. Lines show predicted probabilities that children's first device-relevant reach was directed toward the partner's on-screen face or the webcam. Shaded bands indicate 95% confidence intervals from parametric simulation. Older children were increasingly likely to initiate webcam-directed responding, whereas younger children were more likely to begin by acting on the partner's face.

To account for the sequential structure of the task, in which children progressed to later scaffolding levels only after failing earlier stages, we modeled children's first successful webcam blocking using a discrete-time hazard model. This approach estimated the probability of success at each stage, conditional on not having succeeded previously. Across stages, success rates

increased from 5.8% in the Spontaneous condition (5/86) to 11.3% in Emulation (8/71) and 25.0% in Imitation (14/56). The logistic hazard model revealed a significant effect of age: older children were more likely to achieve their first successful webcam-blocking response, $b = 1.25$, $SE = 0.43$, $z = 2.91$, $p = .004$, OR = 3.49, 95% CI [1.50, 8.11].

The probability of success varied across scaffolding stages. Relative to the Spontaneous condition, the Emulation stage did not significantly increase the likelihood of success, $b = 0.81$, $SE = 0.60$, $z = 1.34$, $p = .179$, OR = 2.25, 95% CI [0.69, 7.36]. In contrast, the Imitation stage was associated with a higher likelihood of success, $b = 1.94$, $SE = 0.58$, $z = 3.36$, $p < .001$, OR = 6.92, 95% CI [2.24, 21.44]. These results indicate that children's likelihood of adopting a webcam-based strategy increased with age and was further facilitated at the highest level of scaffolding.

To test whether the effect of scaffolding varied with age, we compared models with and without an Age × Stage interaction. Adding the interaction did not significantly improve model fit, likelihood-ratio $\chi^2(2) = 1.65$, $p = .44$. Thus, although older children were more likely to succeed overall, the relative benefit of scaffolding did not differ reliably by age. See Figure 1.7.

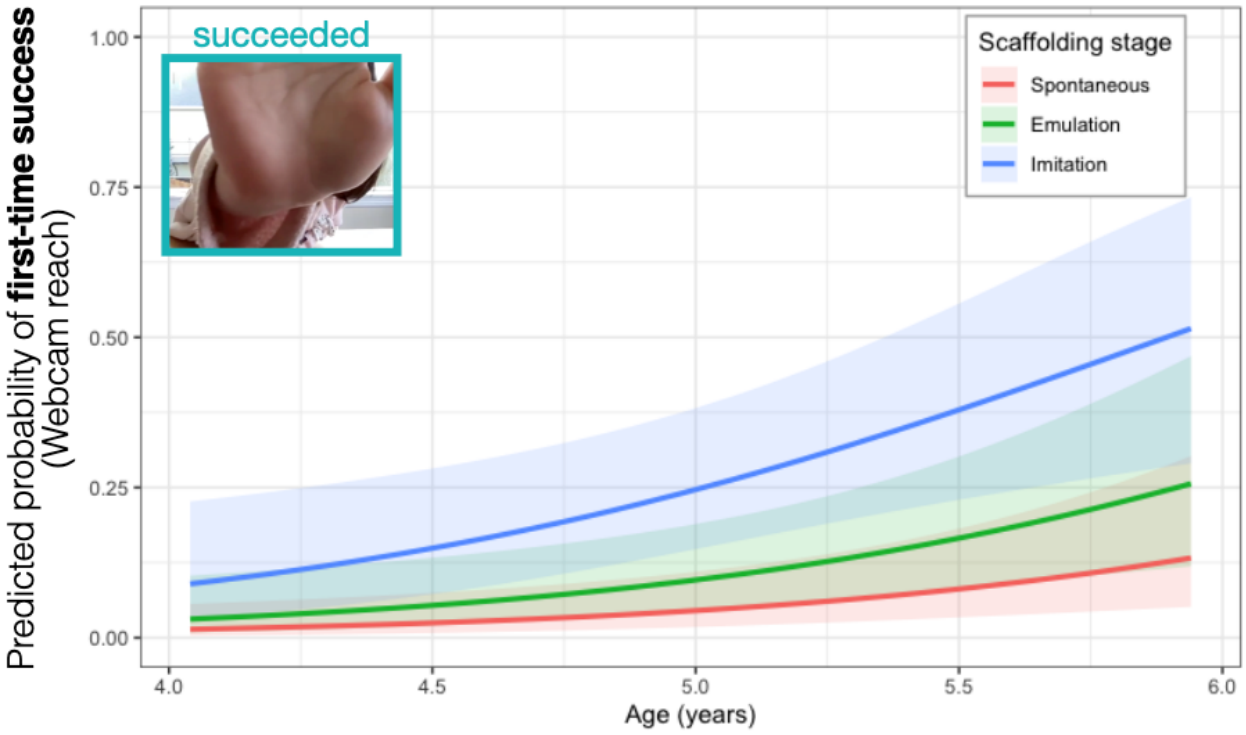

Figure 1.7: Model-predicted probability of first successful webcam blocking across age and scaffolding stages. Lines show predicted probabilities from a discrete-time hazard model, with shaded bands indicating 95% confidence intervals. Because children progressed to later stages only after failing earlier ones, predictions reflect success probabilities conditional on prior failure. Later stages show higher predicted probabilities of success, particularly at older ages, with the largest increase observed in the Imitation stage.

To examine how children's behavioral strategies evolved across stages, we fit a multinomial logistic model predicting later reach target from earlier reach target, controlling for transition stage. Earlier reach behavior significantly improved model fit beyond the transition stage alone, likelihood-ratio $\chi^2(6) = 38.4$, $p < .001$, indicating that children's subsequent actions depended on their preceding behavioral strategy. Model-predicted transition probabilities showed substantial within each reach target and a selective pathway from self-view–directed to webcam-directed responses. Direct transitions from partner-face–directed responses to the webcam-directed responses were rare. See Figure 1.8.

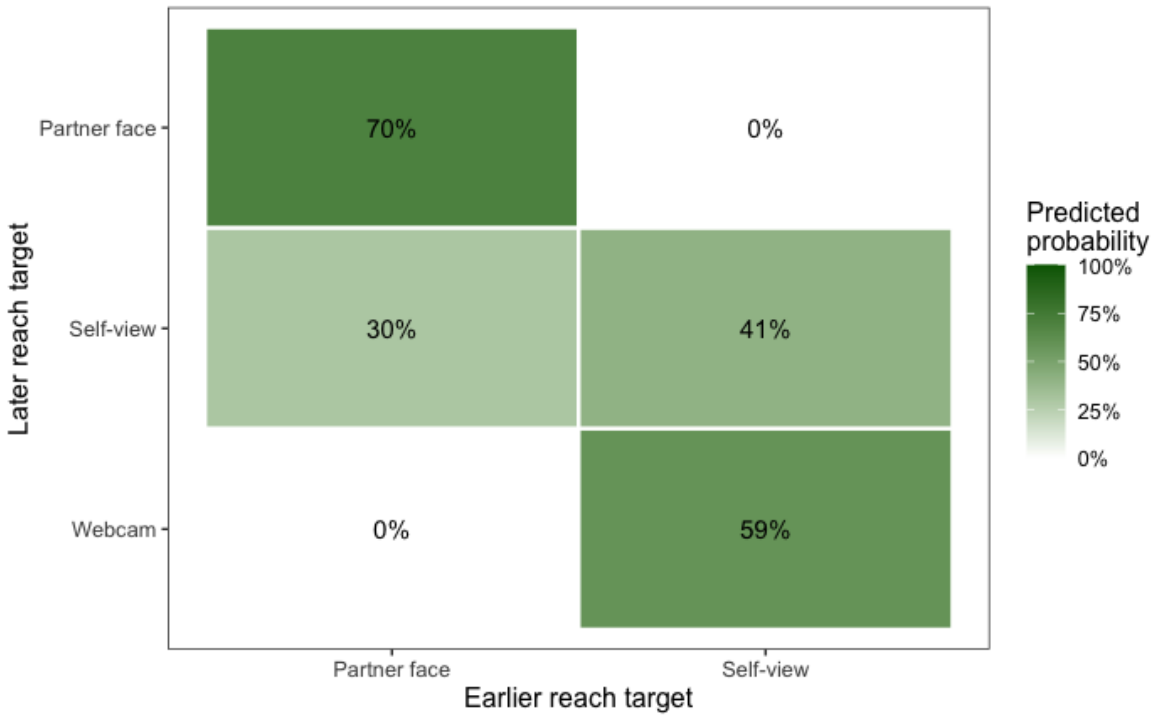

Figure 1.8: Model-predicted transition probabilities among device-relevant reach strategies. Predicted probabilities are derived from a multinomial logistic model controlling for age and transition stage. Visualization is restricted to meaningful behavioral strategies (partner face, self-view, webcam), with probabilities renormalized after excluding task-irrelevant responses. Webcam-directed responses serve as a terminal state and are therefore not included as sources of transitions.

We next tested whether transition patterns varied with age. Adding age significantly improved model fit, likelihood-ratio $\chi^2(3) = 14.80$, $p = .002$, indicating that older children differed in their overall response tendencies. However, the interaction between earlier reach target and age was not significant, likelihood-ratio $\chi^2(6) = 1.16$, $p = .979$. Thus, the association between earlier and later reach targets did not vary reliably with age. The probability of transitioning from self-view-directed to webcam-directed responses also did not vary reliably with age or scaffolding stages. Instead, developmental change was reflected in an overall increase in webcam-directed responses: older children were more likely to reach toward the webcam regardless of their preceding strategy.

Experiment 3.

A mixed-effects logistic regression revealed a strong effect of question type on response side, likelihood-ratio $\chi^2(2) = 238.41$, $p < .001$, with no main effect of age, likelihood-ratio $\chi^2(1) = 0.37$, $p = .54$. Children were highly likely to respond on the self-view side in the See condition ($M = 87\%$) and See-only condition ($M = 79\%$), but predominantly responded on the partner-view side in the Touch condition ($M = 19\%$). Pairwise comparisons confirmed that all three conditions differed significantly from one another (all ps $< .05$). This pattern indicates that children adjusted their interpretation of self-view information to the informational demands of the task: they treated it as relevant to visual access, but not to action-based judgments. See Figure 1.10.

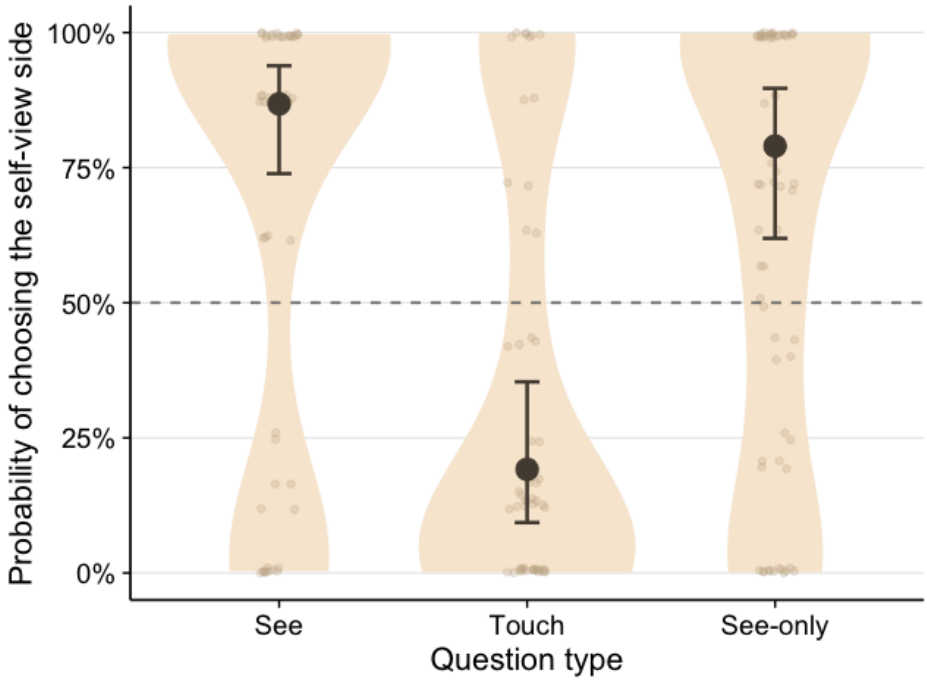

Figure 1.9: Probability of choosing the self-view side by question type. Violin plots show the distribution of participant-level proportions of self-view responses within each condition. Gray points represent individual participants. Black points with error bars indicate model-predicted probabilities and 95% confidence intervals from a mixed-effects logistic regression with question type and age as fixed effects and participant as a random intercept. The dashed line marks chance-level responding (50%).

Discussion

The present study investigated how children aged 4.0 to 5.9 years infer others' visual perspective in video chat, a context in which perceptual input (self-view vs. partners face on the screen) provide challenges but also opportunities to learn video chat partner's visual access. Across three preregistered experiments, we show that children progressively reorganize how they infer visual access by using the self-view as a source of information about the video chat partner's perspective. This shift is evident across levels of cognition: in action (Experiment 1), in learning under scaffolding (Experiment 2), and in explicit reasoning (Experiment 3). Overall, these findings demonstrate that children can move beyond a default, gaze-based understanding of vision and construct a representation-based account of visual access that is adapted to the constraints of video-mediated communication.

What is the mechanism underlying this shift? In face-to-face interaction, children can rely on a robust heuristic that visual access is grounded in others' eyes (Baron-Cohen, 1995; Langton & Bruce, 1999). In video chat, however, this heuristic is misleading, as visual access depends on an unseen device (i.e., a webcam). In Experiment 1, when the self-view is unavailable, children overwhelmingly direct objects toward the partner's on-screen face, consistent with a gaze-based strategy. This pattern reflects a principled reliance on previously valid assumptions about how vision works. When the self-view is available, they increasingly direct objects toward the webcam, indicating a shift to a representation-based inference. Specifically, children resolve this mismatch by re-grounding perspective-taking in a new representational cue: they learn to interpret the self-view as a depiction of what their partner can see and use this information to guide both action and judgment.

In particular, the self-view does not function merely as an attentional cue. Although sensitivity to gaze and faces emerges early and strongly guides attention (Langton et al., 2000; Capparini et al., 2023), attention alone cannot account for children's success on the tasks. While gaze allocation to the self-view is associated with successful behavior, this relationship is largely explained by age, suggesting that attentional preference alone is insufficient to drive correct responses. Instead, successful task performance depends on how children interpret the informational content of the self-view. This interpretation is selective and context sensitive. In Experiment 3, children preferentially relied on the self-view when reasoning about what the partner can see, but not when reasoning about what the partner can touch, indicating that they treat the self-view as relevant for visual access but not for action-based judgments. Together, these findings suggest that children actively infer the representational meaning of perceptual input, rather than simply responding to its salience.

The developmental pattern observed in Experiments 1 and 2 provides further insight into the processes supporting this inference. In Experiment 1, the probability of correct webcam-directed action increased with age, reaching chance levels around 4.95 years and rising thereafter. This developmental trajectory aligns with prior work showing that flexible perspective-taking and theory of mind reasoning become more robust between approximately 4 and 5 years of age (Flavell et al., 1981; Moll & Meltzoff, 2011; Wellman, 2014). In Experiment 2, older children were more likely to initiate and adopt webcam-based strategies, and were more likely to succeed in identifying the correct source of visual access. However, the structure of behavioral transitions remained stable across age, suggesting that developmental change reflects an increasing likelihood of adopting a particular strategy rather than a fundamental reorganization of the learning pathway.

By contrast, Experiment 3 revealed that even younger children (4-year-olds) can explicitly use the self-view to reason about visual access, with no reliable effect of age on performance. This apparent dissociation between action and explicit reasoning is consistent with prior findings showing that children can succeed in verbal tasks before they can consistently apply the same knowledge in action (DeLoache, 2002; Wellman, 2014). We propose that children acquire the representational insight that the self-view reflects another's perspective relatively early, but that applying this insight in real-time coordination requires additional development in action selection, inhibitory control, and integration of perceptual information. This account helps explain why children may demonstrate competence in explicit reasoning tasks while continuing to rely on suboptimal strategies in interactive contexts.

The learning dynamics observed in Experiment 2 further clarify how children transition between strategies. Scaffolding revealed that success increased most strongly at the imitation stage, whereas observing outcomes alone (emulation) was insufficient to reliably shift behavior. This pattern is consistent with prior work showing that imitation provides more direct support for action mapping than emulation, particularly when causal structure is not yet understood (Tomasello et al., 2005). Moreover, children's actions showed strong persistence, with subsequent behavior depending on prior strategy use. These findings indicate that learning proceeds through structured behavioral pathways rather than through undirected exploration. The self-view appears to function as an intermediate representation that bridges incorrect and correct strategies, supporting a gradual re-mapping of visual access from on-screen cues to the webcam. This process highlights the importance of action-based engagement in learning how perceptual representations map onto underlying causal structures (Gibson, 1979; Smith & Gasser, 2005).

These findings bear directly on theoretical accounts of visual perspective-taking. Modularity accounts propose that perspective-taking relies on an automatic system anchored in gaze and line of sight (Baron-Cohen, 1995; Leslie, 1987, 1994), which should be difficult to override in contexts where these cues are misleading. In contrast, constructivist accounts propose that children actively revise their understanding of others' minds based on experience (Gopnik & Wellman, 2012; Wellman, 2014). The present results provide strong support for the latter view. Children do not remain bound to gaze-based cues; instead, they flexibly adapt their inferences when provided with an alternative source of information. However, this flexibility is constrained by developmental factors, including the ability to interpret representations, suppress salient but misleading cues, and coordinate perception with action.

More broadly, the present findings speak to how humans interpret perceptual information in technologically mediated environments. In the physical world, perception is typically treated as a direct guide to reality. In video chat, however, perceptual input is filtered and transformed by representational systems, such that its relationship to underlying states must be inferred. Our results suggest that children approach this problem by using available representations, such as the self-view, as tools for aligning perception with reality. In this sense, the self-view functions not simply as a display feature, but as a resource for constructing an internal model of how perception operates in a mediated system. This perspective is consistent with accounts of perception as an active, inferential process shaped by interaction with the environment (Gibson, 1979; Varela et al., 1991), and with recent work emphasizing the role of experience and inference in theory of mind development (Gopnik & Wellman, 2012).

Children's prior experience with video chat did not strongly predict their use of the self-view. Neither frequency nor duration of use, nor device or platform type, reliably accounted for

variation in behavior. Instead, only the age at first exposure showed a modest association, with earlier exposure linked to greater use of the self-view. This pattern suggests that experience alone is insufficient to drive the observed changes. Rather, children's ability to benefit from experience depends on their developmental readiness to interpret and use representational information. Exposure may provide opportunities for learning, but developmental constraints determine how that input is understood (Wellman, 2014).

Several limitations of the present work should be noted. First, the study was conducted in a controlled video chat setup with standardized interfaces, which may not capture the full variability of real-world interactions. Second, the partner's responses were held constant, limiting the role of contingent social feedback in driving learning. Third, while children learned to use the self-view to guide behavior, many did not demonstrate a full understanding of the webcam as the source of visual input, suggesting a partial dissociation between using a representation and understanding the underlying mechanism. This dissociation echoes broader findings that children can use representations effectively without fully understanding their causal basis (DeLoache, 2002).

In conclusion, the present study shows that children are not passively confused by video-mediated interaction. Instead, they actively construct new mappings between perceptual representations and underlying reality. By learning to use the self-view as a source of information about another person's perspective, children adapt core processes of social cognition to a novel technological context. These findings highlight the flexibility of perspective-taking and underscore the importance of understanding how humans interpret and act on representations in an increasingly mediated world.

Methods

Approval for the study was obtained from the Institutional Review Board at the University of California, San Diego (IRB protocol #161173).

In all experiments, children participated in the study from their home via Zoom on a computer with a mounted webcam. During the Zoom setup phase in each experiment, parents were instructed to set Zoom to full-screen mode and enable side-by-side display of the video feeds regardless of the condition. Then, if the condition assignment is Self-view off, parents disabled the child's self-view, leaving only the partner-view visible. In the Self-view On condition, children's Self-view position (left vs. right of the partner-view) was counterbalanced across participants in the Self-view On condition. In the Self-view Off condition, the self-view was hidden, leaving only the partner-view visible.

Once the Zoom setup was complete and parental consent for video recording was obtained, the experimenter initiated recording using Zoom's built-in functionality. Children first completed a gaze calibration task to ensure that gaze direction partner-view and self-view (when present) could be reliably identified from the recording.

Participants

Experiment 1 included 136 children (68 4-year-olds and 68 5-year-olds; 68 per condition; 34 per age group). Experiment 2 included 80 children (40 4-year-olds and 40 5-year-olds). Experiment 3 also included 80 children (40 4-year-olds and 40 5-year-olds).

Experiment 1

Design and procedure. Children were randomly assigned within age range (4–5 years vs. 5–6 years) to one of two conditions (between-subjects design): Self-view On or Self-view Off. The experimental phase was identical across conditions. Children completed a show-and-tell task in which they were asked to show four palm-sized toys to the experimenter (i.e., the video chat

partner); one toy per trial, and four trials in total. On each trial, the experimenter prompted, "Where is your toy?" and, if needed, followed up with, "Can I see it closer?" until the child clearly reached toward either the webcam or the screen. If a child did not engage in task-related behavior after three prompts (e.g., discussing the toy without attempting to show it), the experimenter proceeded to the next trial.

At the end of the experiment, children completed a sanity check task assessing awareness of the webcam location. The experimenter asked three questions: "Do you know where the camera is?", "Can you point to where it is?", and "Can you show me your last toy again?"

Following the experiment, parents completed a brief interview about their observations during the session (e.g., whether the child showed objects toward the screen or webcam), whether the child had previously attempted to show objects to a video chat partner, and, if applicable, how they had intervened in the process.

Measures.

(1) Children's reach location. This refers to whether children attempted to hold the toy to the webcam (object visible to the partner), the self-view (object held near the self-view but out of the camera's view), the partner-view (object held near the partner's on-screen face but out of the camera's view), or other locations, in response to the task prompts. If children adjusted their reach location during a trial, only the final reach was used for this measure.

(2) Children's visual reference. This refers to whether children looked toward the webcam (if the child looks at the webcam), self-view (if the child looks at the self-view), partner-view (if the child looked at the partner-view), and other (if the child didn't use self-view or partner-view as a visual reference).

Coding. To minimize the noise in children's behavior caused by distractions from social interactions (e.g., children often looked toward the partner-view when the experimenter was giving a prompt), we only coded the video recordings from the moment when a child heard a question (i.e., "Where is your toy?") until the child stopped moving (at the end of their reach), and when the experimenter was not speaking. All video segments were coded frame-by-frame in Datavyu by trained research assistants. All reach and gaze behaviors were independently coded by two trained research assistants, with high inter-rater reliability (ICC > .90).

Experiment 2

Design and procedure. All children participated in the experiment with their self-view available. During the experimental phase, children completed a blocking task in which they were asked to prevent the video chat partner from seeing them. Three tasks were administered in sequence. In the Spontaneous task (control), the experimenter asked, "Can you block me from seeing you, so I can't see you or anything from where you are?" In the Emulation task (results learning), children first closed their eyes while the experimenter covered the webcam out of view. After children reopened their eyes, the experimenter said, "I have blocked you from seeing me. Can you try to block me from seeing you, like that?" In the Imitation task (action learning), children observed the experimenter cover the webcam and were then asked, "See how I did that? Can you try to block me from seeing you, like that?"

For each task, if a child successfully covered the webcam, the experimenter said, "Okay, you can stop blocking me now!" If the child did not respond within approximately 5 seconds or engaged in unrelated behaviors (e.g., talking), the experimenter repeated the prompt. If the child did not respond with task-related behavior after three prompts, the experimenter proceeded to the

next task. Remaining tasks were skipped once the child successfully covered the webcam during any task.

At the end of the experiment, children completed a sanity check task assessing awareness of the webcam location. The experimenter asked three questions: "Do you know where the camera is?", "Can you point to where it is?", and "Can you try to block me from seeing you again?"

Measures.

(1) Children's reach location. This refers to whether children attempted to cover the webcam, the self-view, the partner-view, or other locations in response to the task prompts. If children adjusted their reach location during a trial, only the final reach was used for this measure.

(2) Children's visual reference. Children's gaze was coded into four categories as in Experiment 1.

Coding. As in Experiment 1, we only coded the video recordings from the moment when a child heard a question (e.g., "Can you block me from seeing you, so I can't see you or anything from where you are?") until the child stopped moving following an action, successfully covered the webcam, and when the experimenter was not speaking. All video segments were coded frame-by-frame in Datavyu by trained research assistants. All sessions were independently coded by two trained research assistants, with high inter-rater reliability (ICC > .90).

Experiment 3

Design and procedure. All children participated in the experiment with their self-view available. Before the experiment, parents were instructed to draw eight sets of shapes on eight pieces of paper approximately the size of sticky notes. The experimenter prepared matching sets

of shapes. These shape sets were used in a shape-naming task, where children identified shapes visible to the video chat partner (i.e., moon, diamond, heart, triangle, star, circle, flower, square) and in a shape-counting task, where children judged the number of shapes visible to the partner (i.e., one, two, or three circles or squares).

During the experimental phase, each child completed a shape-naming task and a shape-counting task; 4 trials for each task type, and 8 in total. The order of task types was counterbalanced across participants. On each trial, both the child and experimenter attached one paper to their foreheads. The paper the child and the experimenter used was matched and swapped across trials. For example, in one trial, the child used paper A, and the experimenter used paper B; in another, the child used paper B, and the experimenter used paper A.

Within each trial, children responded to three question types: a See question, a See-only question, and a Touch question (control condition). Two question orders were counterbalanced across participants but remained fixed within participants: (1) See, Touch, and See-only, and (2) Touch, See, and See-only. The See-only question was always asked last due to its complex syntax structure (involving length and the conjunction "but"), which could challenge children's comprehension.

In the See question, children were asked what the experimenter could see (e.g., "What shapes can I see?" or "How many squares/circles can I see?"). In the Touch question (control), children were asked what the experimenter could touch (e.g., "What shapes could I touch?" or "How many squares/circles could I touch?"). In the See-only question, children were asked what the experimenter could see but not touch (e.g., "What shape can I see, but not touch?" or "How many squares/circles can I see, but could not touch?").

Measures.

(1) Children's answers. This refers to whether children named (or equivalently, pointed to) pictures visible in the self-view, the partner-view, or both in response to the task questions. If children adjusted their reach location during a trial, only the final reach was used for this measure.

(2) Children's visual reference. Children's gaze was coded into four categories as in Experiment 1.

Coding. As in Experiment 1, we only coded the video recordings from the moment when a child heard a question (e.g., "What shapes can I see?") until the child completed their answer, and when the experimenter was not speaking. All video segments were coded frame by frame in Datavyu by trained research assistants. All sessions were independently coded by two trained research assistants, with high inter-rater reliability (ICC > .90).

Acknowledgments

Chapter 1 is currently being prepared for submission for publication: Lin, C., & Schachner, A. (In Prep). From eyes to webcam: Children learn video chat partner's visual access through self-view.

REFERENCES

Baron-Cohen, S. (1997). *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press.
Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, *8*(6), 535–543.

Capparini, C., To, M. P. S., & Reid, V. M. (2023). Should I follow your virtual gaze? Infants' gaze following over video call. In *Journal of Experimental Child Psychology* (Vol. 226, p. 105554). https://doi.org/10.1016/j.jecp.2022.105554

DeLoache, J. S. (2007). Early development of the understanding and use of symbolic artifacts. In *Blackwell Handbook of Childhood Cognitive Development* (pp. 206–226). Blackwell Publishers Ltd.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1-Level 2 distinction. *Developmental Psychology*, *17*(1), 99–103.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108.

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). *Animal Cognition*, *8*(3), 164–181.

Langton, S. R. H., & Bruce, V. (1999). Reflexive Visual Orienting in Response to the Social Attention of Others. *Visual Cognition*, *6*(5), 541–567.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review*, *94*(4), 412–426.

Leslie, A. M. (1994). ToMM, ToBY, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind* (pp. 119–148). Cambridge University Press.

McClure, E. R., Chentsova-Dutton, Y. E., Holochwost, S. J., Parrott, W. G., & Barr, R. (2018). Look At That! Video Chat and Joint Visual Attention Development Among Babies and Toddlers. *Child Development*, *89*(1), 27–36.

Moll, H., & Meltzoff, A. (2011a). Perspective-Taking and its Foundation in Joint Attention. In *Perception, Causation, and Objectivity* (pp. 286–304). https://doi.org/10.1093/acprof:oso/9780199692040.003.0016

Moll, H., & Meltzoff, A. N. (2011b). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development*, *82*(2), 661–673.

Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2014). Skype me! Socially contingent interactions help toddlers learn language. *Child Development*, *85*(3), 956–970.

Rosengren, K. S., Kirkorian, H., Choi, K., Jiang, M. J., Raimer, C., Tolkin, E., & Sartin-Tarm, A. (2021). Attempting to break the fourth wall: Young children's action errors with screen media. *Human Behavior and Emerging Technologies*, *hbe2.273*. https://doi.org/10.1002/hbe2.273

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *The Behavioral and Brain Sciences*, *16*(3), 495–511.

Troseth, G. L., Saylor, M. M., & Archer, A. H. (2006). Young children's use of video as a source of socially relevant information. *Child Development*, *77*(3), 786–799.

Wellman, H. M. (2014). *Making Minds: How Theory of Mind Develops*. Oxford University Press.

CHAPTER 2 PERCEPTUAL CONFLICT IN AUGMENTED REALITY PROMOTES EXPLORATION AND SUPPORTS REALITY JUDGMENTS IN YOUNG USERS

1. Introduction

Augmented reality (AR) superimposes digital content, such as images, text, or animations, onto the physical environment, creating perceptual experiences in which what people *see* does not always match what is *physically* there. As AR content becomes more realistic and seamlessly integrated into children's daily life through educational apps, interactive storybooks, and social media filters, children must learn to navigate the complex interplay between appearance and reality. Understanding how children judge the reality status of AR content is critical. If children overattribute physical existence to virtual overlays, they may develop persistent misconceptions about the material world (DeLoache, 2000; Flavell, 1986a) or engage in unsafe behaviors, such as reaching for non-physical objects (Slater & Sanchez-Vives, 2016). Conversely, if children under-explore or prematurely dismiss AR cues, they may miss opportunities for learning that support conceptual development (Barsalou, 2008).

Previous work found that 3-year-olds may treat televised characters as socially or physically co-present rather than symbolic representations (Troseth et al. 2006). AR may intensify this challenge by adding spatial alignment and temporal synchrony, increasing perceptual realism while weakening cues that signal content as symbolic. For instance, although children aged 5-6 years engaged with AR-enhanced materials, their learning tended to remain at a perceptual level, with children primarily describing visible features of virtual objects rather than generating deeper, inferential interpretations (Yilmaz 2016). These findings indicate that AR's perceptual richness may support attention and engagement while simultaneously constraining symbolic interpretation, potentially reinforcing a focus on surface features over

abstract meaning. Investigating the cognitive mechanisms underlying children's responses to AR content is therefore essential for understanding children's interaction with symbolic content, and for designing learning technologies.

When do children decide to accept AR content passively, versus explore its relation to the physical world? How does exploration affect their ontological status judgment of AR content? We hypothesize that perceptual conflict in AR (i.e., a misalignment of visual cues and expectations) serves as a cognitive "spotlight" that draws attention, creates uncertainty and invites exploration. Thereby, AR creates novel contexts that elicit children's reasoning, transforming perceptual engagement into epistemic learning about AR content through exploration to test reality and update beliefs. The present study investigates how children detect and respond to AR objects under conditions that either present or do not present perceptual conflicts and whether violations of visual expectations trigger error-driven exploration and learning.

1.1 Perceptual conflict in AR as a cue for exploration

Perceptual conflicts can trigger deeper processing. Prediction-error accounts propose that a mismatch between expected and observed outcomes motivates children to revise their mental models (Gopnik, 2012). This form of error-driven learning is evident in preschoolers, who increase exploratory play when faced with ambiguous or conflicting evidence (Lapidow et al., 2022; Schulz & Bonawitz, 2007). Piaget's notion of cognitive disequilibrium similarly holds that conflicts prompt accommodation to resolve inconsistencies (Piaget, 1952). In cognitive science, conflict-monitoring models highlight the brain's detection of incompatible inputs, which increases attentional engagement to resolve ambiguity (Botvinick et al., 2001). These accounts

converge on the idea that when "something seems off," children often attempt to investigate why.

AR can easily lead to predictive errors in young children. From infancy, children are sensitive to the correspondence between their actions and visual feedback, a form of sensorimotor contingency (O'Regan & Noë, 2001; Rochat & Striano, 2000). Consider live video feeds. Multiple displays of the same scene are usually expected to match their counterparts in the physical world; when live feeds fail to align, the mismatch can serve as a meaningful cue that something is amiss (Amsterdam 1972). When an AR filter introduces a discrepancy, for example, showing a digital overlay in one view but not another, this violates children's intuitive expectations and may serve as potent cues for reevaluation, prompting children to shift from passive viewing to investigate the source of the mismatch. This form of error-driven learning is evident in preschoolers, who increase exploratory play when faced with ambiguous or conflicting evidence (Schulz & Bonawitz, 2007).

Nevertheless, other theoretical accounts caution that observing perceptual conflict in AR may not automatically trigger deeper processing if cognitive resources are taxed or the motivation to resolve uncertainty is low. Children have limited working memory resources (Cowan, 2014). Cognitive overload can occur when too many conflicting or novel cues must be processed simultaneously (Sweller, 1994). Empirical syntheses indicate that immersive media such as VR or AR can heighten cognitive load and reduce active engagement, leading to more passive observation (Makransky & Petersen, 2021). Relatedly, when visual input is conceptually disfluent or inconsistent, it may feel aversive or confusing, leading children to minimize engagement (Alter & Oppenheimer, 2009). Therefore, perceptual conflicts in AR, such as objects

that appear real yet fail to obey physical laws, may overwhelm children's capacity for causal reasoning and conflict resolution.

We focus on 3-6-year-olds as an especially informative age group for studying how children respond to AR content. By this period, children have developed an intuitive understanding of physical principles and how physical objects are expected to behave (e.g., solidity, continuity, and causality; Hood et al., 2003; Lewry et al., 2021; Lin et al., 2022). Yet their conceptual understanding of symbolic and representational content is still emerging, which can lead them to interpret perceptually compelling non-real entities as physically real (Tullos & Woolley, 2009; Woolley & E Ghossainy, 2013). Moreover, children in this age range possess the perceptual and motor abilities needed for deliberate, active exploration. Yet they are still developing the capacity to evaluate visually compelling displays critically and to distinguish appearance from underlying reality (Lane & Harris, 2014). This intersection of robust physical expectations, emerging representational insight, and growing exploratory competence creates a critical window in which perceptual conflict in AR may be meaningful, yet challenging, for young children.

1.2 Exploration as a mechanism for epistemic updating

Developmental research characterizes exploration as a form of hypothesis testing, in which children generate actions to evaluate competing explanations for ambiguous outcomes (Gopnik, 2012; Schulz & Bonawitz, 2007). When encountering uncertainty, children calibrate their information-seeking strategies to maximize causal insight (Lapidow et al., 2022). For instance, preschoolers manipulate objects more extensively after witnessing unexpected events and selectively direct their exploration toward the source of the discrepancy (Perez & Feigenson, 2022; Stahl & Feigenson, 2015).

When AR filters introduce visually inconsistent live feeds across screens, children may engage in active exploration to test hypotheses (e.g., "When the real world object moves, the screen content should follow exactly"). Specifically, across the preschool years, exploratory behaviors become increasingly goal-directed and informative. Younger preschoolers (around 4 years) often respond to ambiguity with simpler, object-centered actions, such as inspecting or manipulating objects directly (C. Cook et al., 2011). With age, children's reasoning about unexpected outcomes becomes more sophisticated: 6-year-olds were more likely than 4-year-olds to explain surprising events by appealing to underlying causal mechanisms rather than surface appearance, particularly when those events violated their expectations (Legare et al., 2010). This developmental shift likely supports more deliberate and informative forms of exploration in situations that require evaluating competing explanations, and may be reflected in how children respond to perceptual conflict induced by AR content.

Children often use the evidence they collect to update prior assumptions and construct more accurate causal explanations (Bonawitz et al. 2012; Legare 2012). However, exploration in AR contexts may not always lead children to correct conclusions. When children investigate surprising events, their interpretations of the evidence depend on developing conceptual abilities, including distinguishing between fantasy and reality, and understanding what is possible and what is impossible in the physical world (Flavell 1986; Shtulman and Carey 2007). For instance, preschoolers often show difficulty evaluating the feasibility of novel or ambiguous events, sometimes treating fantastical or representational entities as potentially real when perceptual cues are compelling (Tullos & Woolley, 2009; Woolley & E Ghossainy, 2013). Accordingly, a child who explores conflicting displays may conclude that the AR transformation reflects a real, unusual physical event rather than a representational overlay.

Therefore, we hypothesize that exploration serves as a key mechanism linking children's observation of perceptual conflict to reasoning, but its effectiveness depends on the conceptual resources children bring to the task. When children notice a discrepancy and possess a sufficiently robust understanding of reality, causality, and possibility, exploration can help them generate informative evidence and revise mistaken impressions. Alternatively, without such understanding, children may fail to explore; or exploration may lead to misinterpretation, reinforcing perceptually compelling but incorrect conclusions.

1.3 The current study

The current study investigates how children aged 3 to 6 judge the ontological status of screen content (authentic vs. AR mediated) and how their explorations and judgement are affected by perceptual conflict induced by AR (i.e., discrepancies between authentic versus AR-altered feeds).

Across two experiments, we employed a dual-screen setup of two tablets, each streaming a live view of the child from the front-facing camera. Children were asked to identify a picture which had been placed on their forehead: In this position, they could not see the image directly, but could see it reflected in both on-screen video feeds. We experimentally manipulated whether the two video feeds produced authentic or AR mediated content (using an AR filter). We also manipulated whether the video feeds produced perceptually consistent or conflicting appearances. In consistent conditions, both video feeds showed the same image, either because AR was not used for either one or because the same AR filter was used for both. In the conflicting conditions, the child saw two different images on the two different tablets: One used an AR filter to transform the card's image, for example, turning an elephant into a pig (see Figure 2.1).

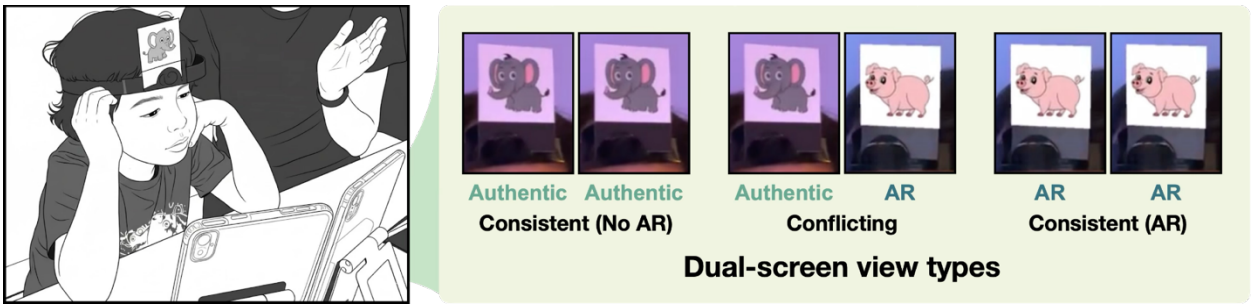

Figure 2.1: Dual-screen paradigm used in both experiments. Two tablets each streamed a live view of the child from the front-facing camera. In this illustration, the card on the child's headband is an elephant. Using an AR filter, we manipulated the presence or absence of perceptual conflict in three view types: (1) Consistent (No AR), where both screens showed identical, authentic images (i.e., elephants in this example), (2) Conflicting, where one screen showed an authentic image and one an AR overlay (i.e., a virtual image of pig replacing the elephant in this example), and (3) Consistent (AR), where both screens displayed identical AR overlays.

Because live feeds of the same scene typically yield perceptually matching information, we predicted that children would explore less when two video feeds appeared consistent, regardless of the ontological types of the screen content (authentic in Experiment 1 or AR in Experiment 2). In contrast, perceptual conflict would create prediction errors, prompting children to explore further to resolve the uncertainty (e.g., inspecting the physical card or testing the system by inducing AR "glitches" on the screen) and supporting more accurate judgments of the authenticity of the screen content. Alternatively, some children might not interpret the discrepancy as meaningful, and fail to explore in this case, leading to acceptance of screen content (in perceptually consistent views) or to arbitrary endorsement of either screen's content as real (in perceptually conflicting views).

Experiment 2 further tested whether perceptual conflict not only triggers immediate exploration but also promotes lasting conceptual change in children's understanding of screen content. To do this, we used a pretest-intervention-posttest design (with perceptual conflict as the intervention). We predicted that exposure to conflicting visual cues would initiate an epistemic

shift, enabling children to reevaluate the ontological status of screen objects more accurately, even after the perceptual conflict is removed.

Both experiments were preregistered, and the preregistrations, along with the study materials, are publicly available at https://osf.io/v2edr/overview (Experiment 1) and https://osf.io/s9wv2/overview (Experiment 2).

2. Experiment 1

Experiment 1 specifically asks, (1) Can children correctly identify the authentic image content when AR creates a perceptual conflict between two views (AR vs. authentic), compared with when video feeds show the same, authentic (non-AR) images? (2) Does perceptual conflict elicit more exploration than a consistent display? And (3) Does exploration predict accurate judgments of the authentic content?

We predicted that perceptually consistent views would align with children's expectation and lead them to judge that screen content displays the authentic physical card content, without actively seeking additional evidence through exploration. In contrast, perceptually conflicting views (e.g., seeing one video feed showing an elephant and the other a pig) would create prediction errors, prompting children to engage in active exploration to resolve the uncertainty, such as removing the card to inspect the image directly, or leaning over to view the card in a traditional mirror. We predict that active exploration will lead to a greater ability to identify the authentic image, distinguishing it from the AR image. If perceptual conflict does not foster reasoning, or if active exploration does not lead to correct reality judgments, then we may instead observe children making arbitrary judgments, choosing randomly between the two options in the perceptually conflicting view. We also explored possible changes with age and

prior AR experience, asking whether with age or prior AR experience children become more able to accurately distinguish AR images as artificial rather than physically real.

2.1 Method

2.1.1 Participants

Participants were N=170 children from 3–6 years of age, inclusive (M = 4.8 years; 75 girls), tested in person at a local museum and preschools in Southern California, USA. This final sample included the preregistered target sample of 3- and 5-year-old children (n = 48 per age group), as well as additional children aged 4, 5 and 6 years who were recruited via convenience sampling at the same locations until the preregistered target numbers for ages 3 and 5 were met. The final sample had relatively balanced representation across age groups: 48 3-year-olds (M_{age} = 3.63, SD = 0.25; 23 girls), 35 4-year-olds (M_{age} = 4.45, SD = 0.31; 15 girls), 57 5-year-olds (M_{age} = 5.50, SD = 0.28; 20 girls), and 30 6-year-olds (M_{age} = 6.53, SD = 0.25; 17 girls).

An additional 175 children were tested but excluded under preregistered exclusion criteria, due to experimenter error (n = 13), parental inference (n = 15), technical failures (n = 40), or failure to pass the warm-up phase or to complete the task (n = 107). Analyses of the preregistered target sample alone (N = 96; 3- and 5-year-olds) yielded qualitatively identical findings.

2.1.2 Stimuli and Materials

Two tablets (iPad minis, 8.3-inch screens) were placed side by side on a table on adjustable-height tablet stands, in portrait orientation (with the front-facing camera at the top). Each tablet displayed the front-facing camera feed in one of two modes: either a non-AR camera showing the unaltered feed (authentic mirror view), or an AR camera applying a real-time AR filter. The AR filter was built using Snap AR - Lens Studio (https://ar.snap.com/lens-studio) and

detected the card image, replacing it with a predetermined alternative image. For example, whenever the AR camera recognized the image of an elephant, it replaced the elephant with a pig (and vice versa). A similarly sized mirror was placed about 0.4 inches away from the tablets, also facing the child.

Children were fitted with a black, adjustable-size plastic headband, which had a card slot in the front centered on the forehead (from the picture-guessing game Hedbanz). On each trial, a card with a cartoon image was placed into the slot. The card images varied in color and shape and formed three pairs; for each pair, one image was on the physical card, and the other was the replacement image supplied by the AR filter (frog/rabbit; pig/elephant; butterfly/banana).

2.1.3 Design

Using a within-subjects design, two conditions (2 trials per condition) were administered in a fixed order: the Consistent condition (non-AR), followed by the Conflicting condition. In the Consistent (non-AR) condition, both tablets used the non-AR camera. In the Conflicting condition, one tablet used the AR camera and the other used the non-AR camera; the AR-camera tablet's position (left vs. right) was counterbalanced across trials. Each trial used a different pair of picture cards; the order of pairs was counterbalanced across trials and conditions. See Figure 2.2.

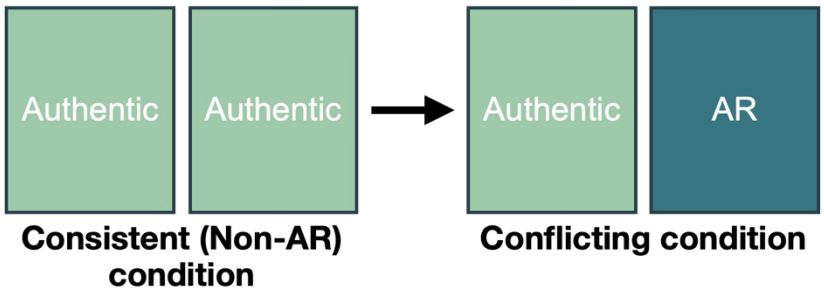

Figure 2.2: Experiment 1 conditions. In the Consistent condition, both screens displayed identical, authentic images, thus lacking perceptual conflict. In the Conflicting condition, the two screens showed conflicting images: one authentic image and one AR overlay.

2.1.4 Procedure

Each child first completed a warm-up task (a picture-guessing game) without tablets to ensure they were comfortable with the experimenter and the task through free exploration. The experimenter randomly drew two picture cards only used in the warm-up phase, and placed them in headbands worn by the child and the experimenter, without showing them to the child (thus, the card in the child's headband remained out of their view). The experimenter then asked, "What is the picture on my forehead?" (which was visible to the child), and then, "Do you know what the picture is on your forehead?" (which was not directly visible to the child). Children who spontaneously explored (by taking the picture off to look at it or by using the mirror) to answer this question then proceeded to the experimental phase. If a child did not spontaneously explore during the game, the experimenter prompted them to do so ("Why not take your picture off and see?" or "Why not use the mirror?"). To minimize reinforcement of exploratory strategies, prompts were used sparingly: each prompt was administered in a separate round, only when children did not follow it, and repeated no more than twice.

Next children completed the experimental phase (two trials of the non-AR condition, followed by two trials of the AR condition). At the beginning of each trial, the experimenter placed a card on the child's forehead (in the headband slot) while covering the card with their hand to ensure that it was kept out of the child's sight. After placing the card, the experimenter said, "Okay!" and moved their hand to uncover the card; at this point the card was not directly visible to the child, but they could see themself on the two tablet screens. The child was then asked, "What is the actual picture on your forehead?" At this point, it was possible for children to spontaneously explore (e.g., by removing the card to look at it or using the mirror), or not, and to answer by either verbalization or pointing to the actual card if they had removed it and it was

in direct view. If the child answered, the experimenter asked, “How do you know?”; after the child’s response, the experimenter then proceeded to the next trial. If the child did not answer, the experimenter repeated the prompt. If the child did not answer after three prompts, the experimenter said, “Okay!” and proceeded to the next trial.

The experimental phase took approximately 5 minutes. An external camera mounted on a tripod recorded the child's face, body language, and verbal responses without capturing the screen information. Both tablets also recorded their screens, capturing the front-facing camera feed along with any AR content presented during the experimental phase.

2.1.5 Measures

Image identification accuracy. To measure children’s accuracy at identifying authentic images (vs. AR images), children’s answers to “What is the actual picture on your forehead?” were categorized as either correct (coded as 1), incorrect (coded as 0, if the answer contains an AR picture or denies any pictures, e.g., “none”), or not applicable (coded as NA, if the child didn’t give an explicit answer, e.g., “I don’t know”). If the child gave multiple answers, their final answer was counted.

Exploratory behaviors. For each trial, children’s exploratory behaviors were coded from the moment the experimenter placed a card on the child’s forehead and said, “Okay!” until the experimenter said, “Okay!” again and proceeded to the next trial. Exploratory behaviors were coded into the following two categories (1 if present, 0 if absent): (1) causal testing: testing whether the screen image follows correspondingly to movements, or instead “glitches” (in the AR), by moving, turning their head, or waving a hand in front of a card’s image while looking at the screens, (2) physical inspection: directly examining the physical card by removing it to look at it, or by examining its reflection in the mirror. For each trial, a single measure of “active

exploration" was coded as 1 if the child engaged in causal testing, physical inspection, or both, and 0 if neither occurred. Additionally, we also coded for the following behaviors: (3) visual exploration of the screen image: providing verbal justification only from visual features shown on the screen (e.g., "it is lighter"; this type of answer was rare, occurring on only 0.7% trials); (4) I don't know, and (5) other, for behaviors that did not fall into other categories.

Prior AR Experience. Parents reported in a questionnaire how often their child usd AR in three contexts: social media (e.g., video filters on Snapchat or Zoom), reading and learning applications (e.g., digital books with AR enhancements), and games (e.g., AR-based gameplay). Responses were collected on a 5-point ordinal frequency scale: 1 = Never, 2 = Yearly, 3 = Monthly, 4 = Weekly, 5 = Every day.

2.1.6 Coding

Verbal responses were transcribed verbatim and exploratory behaviors were coded from external-camera videos in ELAN (Brugman & Russel, 2004); thus coders were naive to correct answers. All coders reached 100% agreement in behavior categorization in a training session. 20% of the videos were annotated by two coders to assess coding reliability, where agreement reached 100% of trials. Ambiguous behaviors in the rest of the video coding were resolved by discussion.

2.1.7 Analysis

Following the preregistered analytic framework, for each dependent variable (e.g., image identification accuracy, active exploration), we compared a series of nested models that incrementally added predictors of interest (e.g., Condition, Age, and their interaction). The preregistered analysis plan specified the use of mixed-effects logistic models. However, some models produced a convergence warning, which can due to relatively few repeated observations

per participant. To ensure robust estimation of population-averaged effects while accounting for within-participant dependence, and to maintain consistency across analyses, we used Generalized Estimating Equations (GEE) with a logit link, which provides a robust approach for modeling correlated binary data with repeated trials within participants. Model fit was evaluated using the Quasi-likelihood under the Independence Criterion (QIC; Pan 2001), with lower values indicating better fit. Differences greater than 2 were interpreted as meaningful improvements following conventional information-criterion guidelines (Burnham and Anderson 2003).

2.2 Results

2.2.1 Image Identification Accuracy

We compared four nested GEE models predicting children's accuracy: (1) intercept-only, (2) Condition-only, (3) Condition + Age, and (4) Condition × Age. The Condition-only model provided the best fit (ΔQIC = 2.54 relative to the next-best model). In this model, children were significantly less accurate in the Conflicting condition than the Consistent (No-AR) condition, b = -2.41, SE = 0.34, Wald $\chi^2(1) = 50.00$, $p < .001$, OR = 0.09, 95% CI [0.05, 0.18]. Neither age nor the Condition × Age interaction improved model fit.

Accuracy was near ceiling in the Consistent (No-AR) condition (M = 95%, 95% CI [92%, 98%]), $t(169) = 27.27$, $p < .001$, and remained above chance in the Conflicting condition (M = 62%, 95% CI [57%, 67%]), $t(169) = 4.48$, $p < .001$; see Figure 2.3.

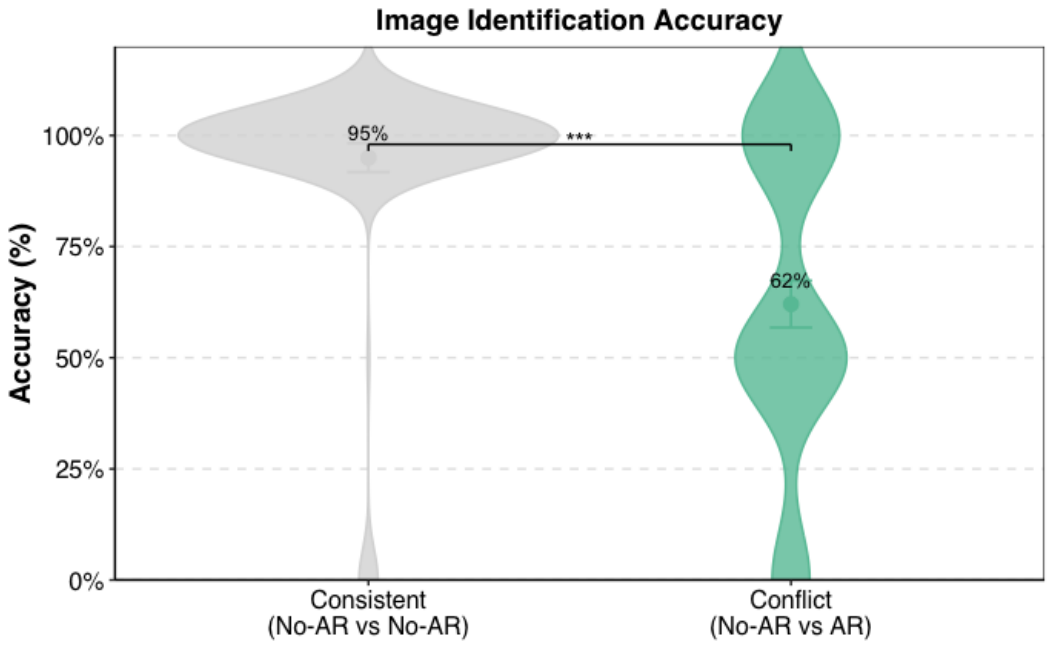

Figure 2.3: Image identification accuracy (accuracy in identifying the real/ non-AR image) across conditions in Experiment 1. Points indicate participant-level means within each condition, overlaid on violin plots showing the distribution of responses. Error bars represent 95% confidence intervals of the mean. Children showed lower accuracy in the Conflicting-AR condition compared to the No-AR condition. Significance annotations reflect the main effect of condition estimated from the GEE model ($p < .05$; $\mathbf{p} < .01$; $\boldsymbol{p} < .001$).

2.2.2 Active Exploration

We compared four nested GEE models predicting children's active exploration behaviors: (1) intercept-only, (2) Condition-only, (3) Condition + Age, and (4) Condition × Age. Model comparison using QIC indicated modest support for a model including a Condition × Age interaction (ΔQIC = 1.8 relative to the next-best additive model). However, the interaction term itself was not statistically significant, $p = .12$, providing limited evidence that the effect of condition varied with age. We therefore interpret the interaction cautiously and focus on the condition effect, which was robust across models. Children explored more in the AR condition than in the No-AR condition, $b = 1.65$, $SE = 0.21$, Wald $\chi^2(1) = 58.87$, $p < .001$, OR = 5.21, 95% CI [3.45, 7.69]. See Figure 2.4.

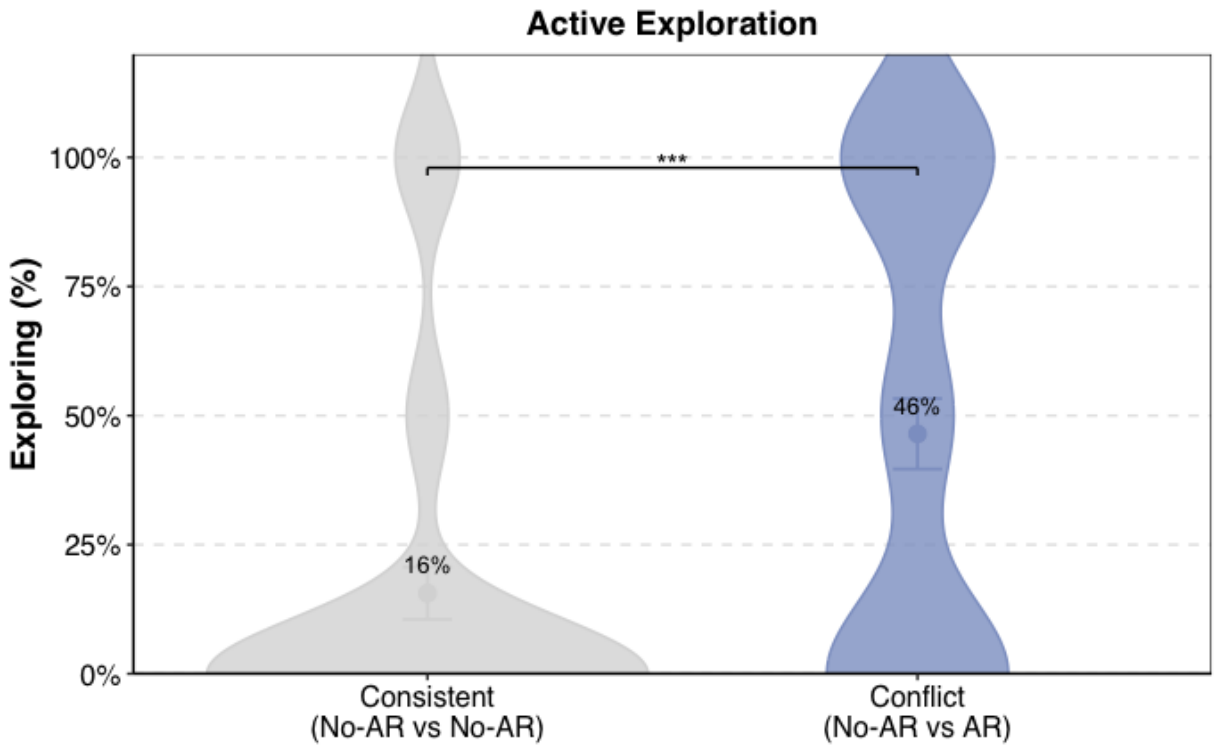

Figure 2.4: Active exploration across conditions in Experiment 1. Points indicate participant-level means within each condition, overlaid on violin plots showing the distribution of responses. Error bars represent 95% confidence intervals of the mean. Children engaged in more active exploration in the Conflicting-AR condition than in the No-AR condition. Significance annotations reflect the main effect of condition estimated from the GEE model ($p < .05$; $\mathbf{p} < .01$; $\boldsymbol{p} < .001$).

We next examined children's use of specific active exploratory strategies, distinguishing between physical inspection and causal testing, and tested whether the use of each strategy differed between the Consistent (No-AR) and Conflicting conditions. Analyses were restricted to children who engaged in active exploration on at least one trial ($n = 95$). For physical inspection, we fit GEE models with a logit link to account for repeated trials within participants, including Condition, Age, and their interaction as predictors. Model comparison using QIC indicated that the Condition-only model provided the best fit (ΔQIC = 4.60 relative to the next-best model), with no evidence that adding Age or the Condition × Age interaction improved model fit. In this model, children were significantly more likely to engage in physical inspection in the Conflicting

condition than in the Consistent (No-AR) condition, $b = 1.75$, $SE = 0.24$, Wald $\chi^2(1) = 52.8$, $p < .001$, OR = 5.73, 95% CI [3.58, 9.18].

We then examined children's use of causal testing using GEEs with the same predictors. Model comparison using QIC indicated that the model including the Condition × Age interaction provided the best fit (ΔQIC = 8.37 relative to the next-best model). In this model, there was a significant main effect of Condition: children were more likely to engage in causal testing in the Conflicting condition than in the Consistent (No-AR) condition, $b = 3.24$, $SE = 0.50$, Wald $\chi^2(1) = 41.46$, $p < .001$, OR = 25.52, 95% CI [9.52, 68.40]. There was also a significant main effect of Age, $b = -1.08$, $SE = 0.37$, Wald $\chi^2(1) = 8.49$, $p = .004$. However, because this main effect was qualified by a significant Condition × Age interaction, $b = 1.09$, $SE = 0.42$, Wald $\chi^2(1) = 6.79$, $p = .009$, the age effect differed by condition. Specifically, the effect of the condition on causal testing became stronger with age.

2.2.3 Active Exploration Predicts Image Identification Accuracy

In line with the preregistered plan, we next asked if children's active exploration in the AR condition predicted their image identification accuracy, by comparing four nested models predicting children's image identification accuracy in the AR condition only: (1) intercept-only, (2) Active Exploration, (3) Active Exploration + Age, and (4) Active Exploration × Age. The model including Active Exploration only provided the best fit to the data (QIC = 441, ΔQIC = 2.12 relative to the next-best model). Children were more likely to correctly identify the object on trials in which they engaged in exploration, $B = 0.52$, $SE = 0.24$, Wald $\chi^2(1) = 4.87$, $p = .027$, OR = 1.69, 95% CI [1.06, 2.68]. Adding age and the exploration × age interaction did not improve model fit, and neither age nor the interaction was significant (both ps > .40). Thus, the

association between active exploration and image-identification accuracy did not vary reliably with age.

2.2.4 Prior AR Experience Does Not Predict Active Exploration

We next asked whether differences in children's active exploration between the No-AR condition and AR conditions were associated with their prior experience with AR. All three AR experience variables were severely right-skewed and showed pronounced floor effects. Social media AR (M = 1.85, SD = 1.09) and Games AR (M = 1.81, SD = 1.22) had low mean frequencies, with over half of participants reporting no use (54.8% and 64.7%, respectively) and very few reporting daily use (1.1% and 3.1%, respectively). Learning AR (M = 2.34, SD = 1.47) showed the highest frequency (48.2%) and variability, with 9.6% reporting daily use. Given the limited variability and pronounced floor effects, responses were dichotomized into "No AR Experience" (scale value =1) versus "Any AR Experience" (scale values 2-5), and then averaged across the three contexts to obtain a single AR experience score for each child. Children's AR experience was not significantly associated with the change in level of active exploration from No-AR to the AR condition, Spearman's $\rho = .09$, $p = .27$, 95% CI [-.07, .24].

2.3 Discussion

Experiment 1 demonstrates that perceptual conflict in AR (two apparently different realities across two camera feeds) drives active exploration, which in turn supports more accurate reality judgments. When children encountered conflicting visual information across screens, they did not passively accept screen information (as they did when the perceptual conflict was absent); instead, they engaged in exploratory actions, such as physically inspecting the card or inducing AR "glitches," that enabled them to evaluate which image corresponded to physical reality.

Crucially, exploration functioned as a mechanism linking perceptual conflict to accurate judgment. Children who engaged in active exploration were more likely to correctly identify the authentic image, indicating that sensorimotor engagement plays a central role in resolving ambiguity about what is real. These results support the hypothesis that perceptual conflict in AR acts as an epistemic cue, signaling that available information may be unreliable and warrant further exploration. Rather than increasing attention alone, perceptual conflict prompted children to generate actions that tested competing interpretations of the visual input.

Notably, these effects did not vary with age between 3 and 6 years, suggesting that the capacity to use exploration to resolve perceptual ambiguity emerges early in development, when task structure supports active inquiry. This finding highlights the role of context in eliciting reality-monitoring processes, rather than reflecting purely age-dependent maturation. Despite robust group-level effects, variability across children was evident. Not all children increased exploration or accuracy in response to conflict, indicating individual differences in responsiveness. These differences were not explained by prior AR experience, suggesting that responsiveness to perceptual conflict may depend more on domain-general factors that differ across individuals, such as curiosity, executive function, or sensitivity to uncertainty.

These findings raise an important question: Is children's exploration driven by AR content more generally, or specifically by the presence of perceptual conflict? A key limitation of Experiment 1 is that perceptual conflict and AR content were confounded: the only condition containing AR also introduced conflict. Thus, it remains unclear whether exploration was driven by AR content itself or specifically by perceptual conflict. Experiment 2 addresses this issue by disentangling these factors and testing whether exploration is driven by perceptual conflict per se or by AR content more broadly. In addition, we examine whether perceptual conflict not only

triggers immediate exploration, but also promotes lasting changes in children's understanding of screen content.

3. Experiment 2

Experiment 2 builds on Experiment 1, primarily investigating two questions: (1) Does perceptual conflict, or AR content alone, prompt children's active exploration? and (2) Does perceptual conflict elicit lasting epistemic updating, affecting children's subsequent understanding of a similar but conflict-free AR scenario?

We hypothesized that perceptually consistent views would align with children's predictions even when the displays are both AR overlays and not authentic; and therefore, that children would accept screen content displays as showing the authentic, physical card without actively seeking additional evidence. Alternatively, if the mere presence of AR is enough to trigger uncertainty, active exploration and learning, children should explore and accurately identify authentic vs. AR content even when the views are perceptually consistent with each other, just as they would for the perceptually conflicting view.

We also hypothesized that exposure to conflicting visual cues would initiate an epistemic shift, enabling children to more accurately assess the reality status of on-screen objects, even after the perceptual conflict is removed. If this is the case, after experiencing a perceptually conflicting view, children would show increased exploratory behavior and improved image identification accuracy even when observing a subsequent perceptually consistent view. Alternatively, perceptual conflict may not elicit even short-term epistemic updating, resulting in no differences in exploratory behavior or judgment accuracy before versus after experiencing a perceptual conflict view.

In addition, we conducted an exploratory analysis to determine how children's prior exposure to different screen media content (e.g., animated vs. live-action, and mixed media with elements of both) moderated their responses to AR displays. In particular, we asked whether children would generalize their experience with mixed media (which combines both animated-and-live-action elements) to AR content, potentially leading them to see AR-like content as common, and making them less motivated to explore.

3.1 Methods

3.1.1 Participants

Participants were N=91 children from 3 to 6 years of age, inclusive (M = 5.08 years; 42 girls), tested in person at the same local museum as Experiment 1. This final sample included the preregistered target sample of 72 children aged 3 to 6 years, N=18 per age group; as well as additional children aged 3 to 6 years who were recruited via convenience sampling at the same location until the registered target numbers for all age groups were met. The final sample had relatively balanced representation across age groups: 20 3-year-olds (M_{age} = 3.58, SD = 0.28; 11 girls), 26 4-year-olds (M_{age} = 4.60, SD = 0.29; 10 girls), 22 5-year-olds (M_{age} = 5.52, SD = 0.32; 10 girls), and 23 6-year-olds (M_{age} = 6.52, SD = 0.25; 11 girls). A total of 38 additional children were tested but excluded under preregistered exclusion criteria, primarily due to experimenter error (n = 20), parental interference (n = 1), technical failures (n = 5), or failure to pass the warm-up phase or to complete the task (n = 12).

3.1.2 Stimuli and Materials

The stimuli and materials were identical to those used in Experiment 1, except that in the conditions without perceptual conflict (consistent conditions), both tablet screens showed the AR

camera, which replaced the real image with a different one using an AR overlay. In the consistent conditions, the same AR image was shown on both tablet screens (see Figure 2.5).

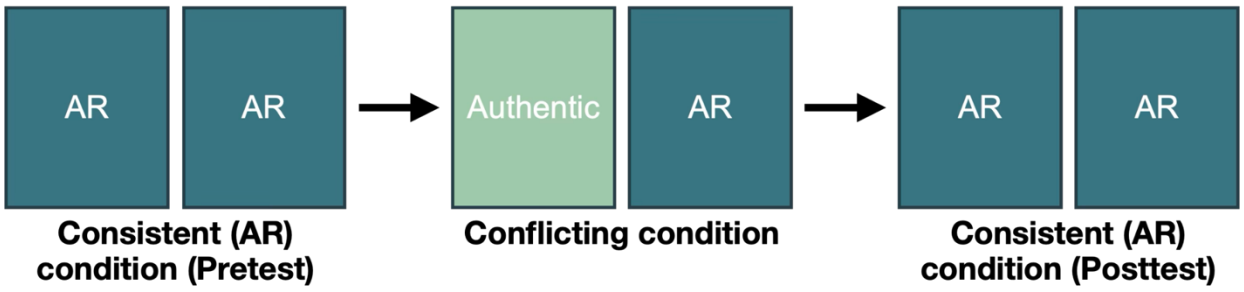

Figure 2.5: Experiment 2 conditions. In the Pretest and Posttest conditions, both screens displayed identical AR overlays. In the Conflicting condition (intervention), one screen showed an authentic image while the other displayed an AR overlay, creating perceptual conflict between the two views.

3.1.3 Design

Using a pretest-intervention-posttest design, three conditions (with two trials per condition) were administered in a fixed order: The Pretest Consistent-AR condition, followed by the Conflicting condition (intervention), and then the Posttest Consistent-AR condition. In both the Pretest and Posttest conditions, both tablets used the AR camera and both screens displayed identical AR overlays. In the Conflicting condition, one tablet used the AR camera and the other the non-AR camera, creating perceptual conflict (as in Experiment 1). The position of the tablet with the AR-camera (left vs. right) was counterbalanced across trials. Each trial used a different pair of picture cards; pair orders were counterbalanced across trials and conditions.

3.1.4 Procedure

The procedure was identical to that of Experiment 1, except for (1) the addition of a third (posttest) consistent condition; and (2) the two consistent conditions showing the AR-overlay on

both tablets (rather than on neither, in Experiment 1). The total duration was approximately 6 minutes.

3.1.5 Measures

Measures were identical to those in Experiment 1, with the exception of the following additions to the parent questionnaire. First, the prior AR experience measure was revised to capture children's AR use more precisely: In addition to the 5-point ordinal scales, parents reported the child's average daily AR duration (in hours or minutes).

Second, we added items to assess children's general screen time and media-type exposure, particularly consumption of animated vs. live action media. Parents reported the child's average daily screen time (hours or minutes) for five categories of device: computer/laptop, tablet/phone, television, video game console, and other devices. Parents also rated the typical composition of children's screen media consumption on a 5-point scale: 1= Almost all is animated, 2 = Mostly animated, with some live-action, 3 = About half animated, half live-action, 4 = Mostly live-action, with some animated, 5 = Almost all is live-action. Parents also reported how often the child watched media that combines animation and live-action (for example, *Mary Poppins, Gabby's Dollhouse*; termed mixed media) on a 5-point frequency scale from "Never" to "Very often."

3.1.6 Coding and Analysis

Coding procedures and analytic approach were identical to those used in Experiment 1. All coders reached 100% agreement in behavior categorization in a training session. 20% of the videos were annotated by two coders to assess coding reliability, where agreement reached 100% of trials. Ambiguous behaviors in the rest of the video coding were resolved by discussion.

3.2 Results

3.2.1 Image Identification Accuracy

In line with the preregistered plan, we compared four nested GEE models with a logit link, predicting children's identification accuracy: (1) intercept-only, (2) Condition-only (Pretest Consistent-AR; Conflicting-AR; Posttest Consistent-AR), (3) Condition + Age, and (4) Condition × Age. The Condition-only model provided the best fit to the data (ΔQIC = 5.23 relative to the next-best model). Adding age and the Condition × Age interaction did not improve model fit, and neither age nor the interaction terms were significant (all *p*s > .20).

Relative to the Pretest Consistent-AR condition, children were more likely to correctly identify the image in the Conflicting-AR condition, $b = 1.43$, $SE = 0.22$, Wald $\chi^2(1) = 41.70$, $p < .001$, OR = 4.19, 95% CI [2.71, 6.48], and in the Posttest condition, $b = 0.50$, $SE = 0.17$, Wald $\chi^2(1) = 8.64$, $p = .003$, OR = 1.65, 95% CI [1.18, 2.30]. Children were also more likely to respond correctly in the Conflicting condition than in the Posttest Consistent-AR condition, $b = 0.94$, SE = 0.21, Wald $\chi^2(1) = 20.00$, $p < .001$, OR = 2.55, 95% CI [1.68, 3.88].

Accuracy was below 50% chance in the Pretest condition ($M = 40\%$, 95% CI [31%, 49%], $t(90) = -2.16$, $p = .033$, and above chance in the Conflicting-AR condition ($M = 74\%$, 95% CI [67%, 81%]), $t(90) = 6.55$, $p < .001$. Accuracy in the Posttest condition did not differ significantly from chance ($M = 52\%$, 95% CI [42%, 62%]), $t(90) = 0.45$, $p = .657$. See Figure 2.6.

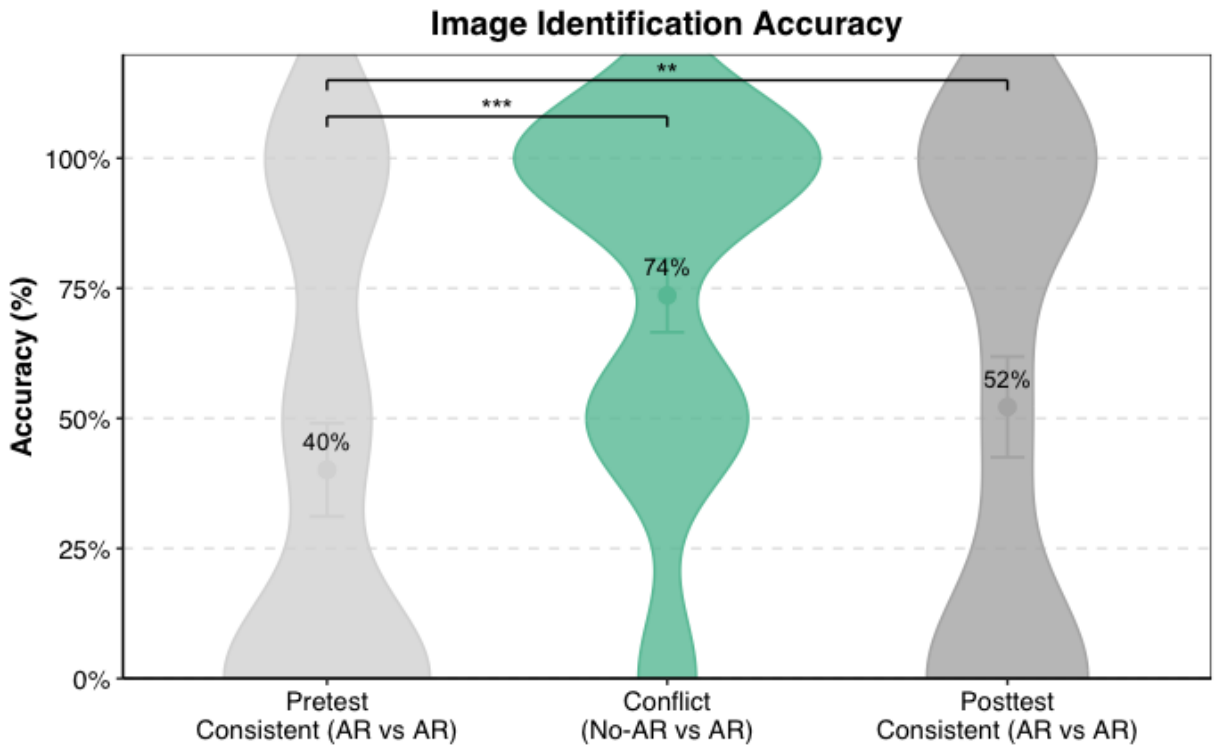

Figure 2.6: Image identification accuracy (accuracy in identifying the real/ non-AR image) across conditions in Experiment 2. Points indicate participant-level means within each condition, overlaid on violin plots showing the distribution of responses. Error bars represent 95% confidence intervals of the mean. Children showed higher accuracy in the Conflicting-AR condition compared to both the Pretest and Posttest conditions, while accuracy in the Posttest condition remained higher than in the Pretest condition. Significance annotations reflect pairwise comparisons estimated from the GEE model using emmeans ($p < .05$; **p** < .01; ***p*** < .001).

3.2.2 Active Exploration: Children Explore More During and After Perceptual Conflict

In line with the preregistered plan, we compared four nested models predicting children's active exploration behaviors: (1) intercept-only, (2) Condition-only, (3) Condition + Age, and (4) Condition × Age. The Condition-only model provided the best fit to the data (ΔQIC = 8.10 relative to the next-best model). Relative to the Pretest condition, children were more likely to engage in exploration in the Conflicting-AR condition, OR = 2.06, and in the Posttest condition, OR = 1.72. Pairwise comparisons confirmed that both the Conflicting and Posttest conditions differed from the Pretest condition (both *p*s < .001), whereas exploration rates did not differ

significantly between the Conflicting and Posttest conditions, OR = 1.20, p = .122. See Figure 2.7.

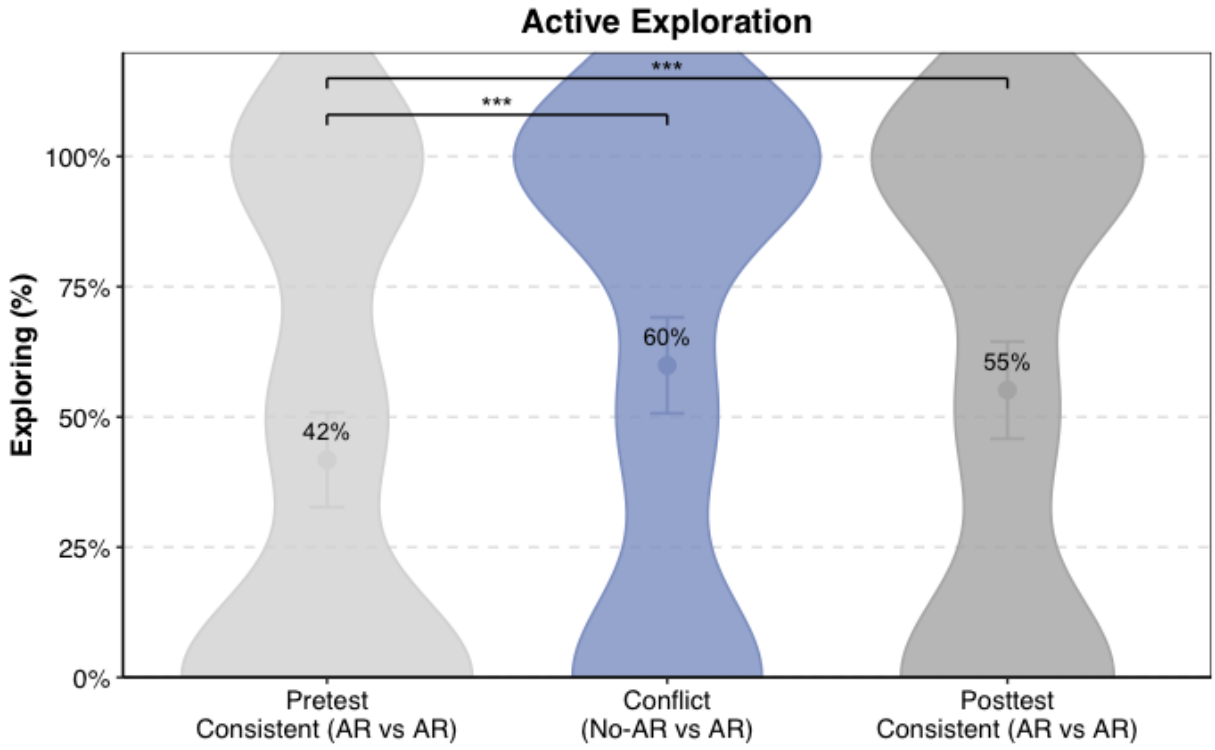

Figure 2.7: Active exploration across conditions in Experiment 2. Points indicate participant-level means within each condition, overlaid on violin plots showing the distribution of responses. Error bars represent 95% confidence intervals of the mean. Children engaged in more active exploration in the Conflicting-AR condition compared to both the Pretest and Posttest conditions, whereas exploration did not significantly differ between the Pretest and Posttest conditions. Significance annotations reflect pairwise comparisons estimated from the GEE model using emmeans ($p < .05$; **p** < .01; ***p*** < .001).

We examined children's use of physical inspection as a means of exploration using GEEs with a logit link, clustering by participant. Analyses were restricted to children who engaged in active exploration on at least one trial. Four nested models were compared: an intercept-only model, a model including Condition, a model including Condition and Age, and a model including the Condition × Age interaction. The model including Condition and Age provided the

best fit to the data (ΔQIC = 2.02 relative to the next-best model). There was also a significant effect of age, such that older children were less likely to engage in physical inspection, b = -0.48, SE = 0.191, Wald $\chi^2(1) = 6.21$, p = .013, OR = 0.62, 95% CI [0.43, 0.90]. Adding the Condition × Age interaction did not improve model fit, and the interaction was not statistically significant (ps > .79).

We then examined children's use of causal testing as a means of exploration using GEEs with the same predictors. Model comparison indicated that the Condition + Age model provided a slightly better fit than the Condition-only model (ΔQIC = 1.40), although the improvement was small. The Condition × Age interaction model did not improve model fit. In the Condition + Age model, children were more likely to engage in causal testing in the Conflicting condition than in the reference condition, b = 0.59, SE = 0.22, Wald $\chi^2(1) = 7.52$, p = .006, OR = 1.81, 95% CI [1.18, 2.77]. The effect of Age was not statistically significant, p = .10.

We conducted a GEE analysis with a logit link to examine developmental changes in children's use of exploration strategies, modeling strategy type (physical inspection vs. causal testing), age (centered at 5 years), and their interaction, with clustering by participant. There was a significant Strategy type × Age interaction, b = 0.53, SE = 0.19, Wald $\chi^2(1) = 7.86$, p = .005, OR = 1.70, 95% CI [1.17, 2.46], indicating an age-related shift in strategy use. Model-based estimates indicated that causal testing became more prevalent than physical inspection at approximately 4.49 years of age, 95% CI [3.56, 5.26]. See Figure 2.8.

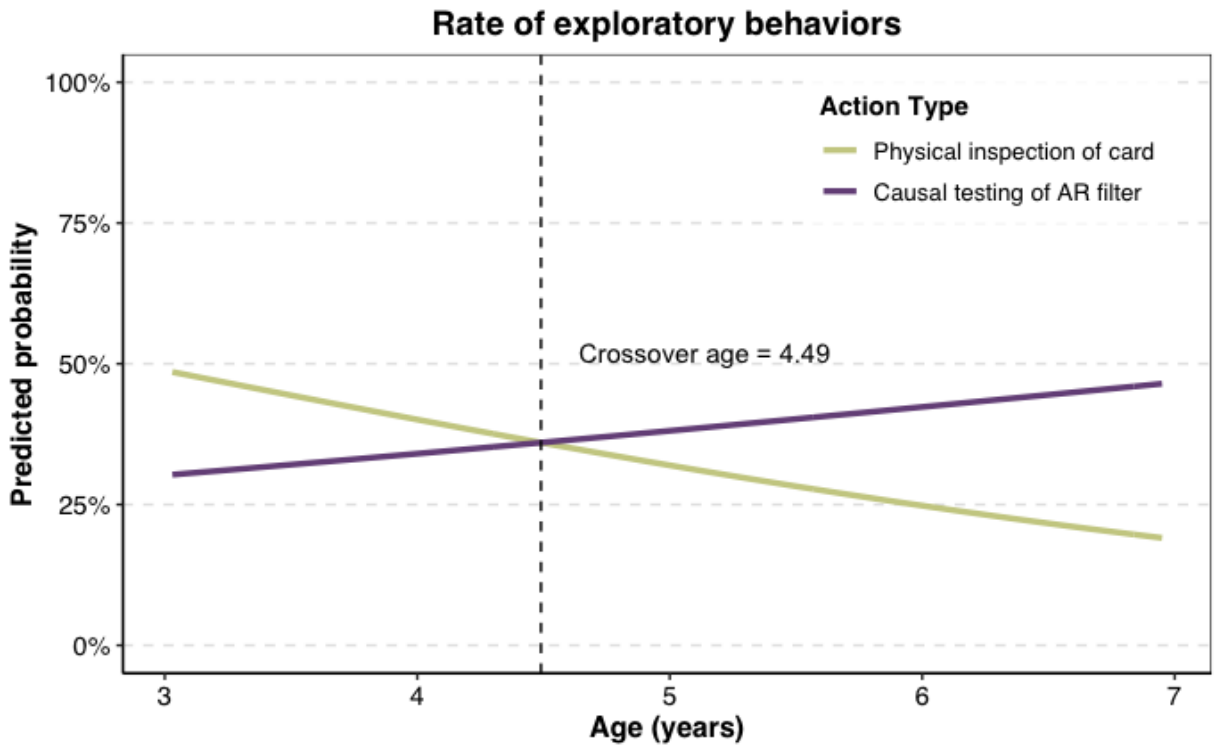

Figure 2.8: Predicted probability of the two active exploratory behaviors as a function of age. Younger children were more likely to engage in physical inspection of the card, while older children increasingly performed causal testing of the AR filter (moving in ways that induce AR "glitches"). The crossover point (at an age of 4.49 years) marks a developmental shift in preferred exploration strategy.

3.2.3 Active Exploration Predicts Image Identification Accuracy

We next asked if children's active exploration predicted their image identification accuracy (including data from all 3 conditions, since AR was present in all conditions). Four nested models were compared, in line with the preregistered plan: (1) intercept-only, (2) Active Exploration-only, (3) Active Exploration + Age, and (4) Active Exploration × Age. The model including exploration provided the best fit to the data (ΔQIC = 2.38 relative to the next-best model). Children were more significantly likely to correctly identify the object on trials in which they engaged in exploration, $b = 2.73$, $SE = 0.28$, Wald $\chi^2(1) = 93.7$, $p < .001$, OR = 15.3, 95% CI [8.79, 26.5]. Adding Age and the Active Exploration × Age interaction did not improve

model fit (ΔQICs $\geq$ 2.38), and neither term was statistically significant (both *p*s > .38). Thus, the association between active exploration and image-identification accuracy did not vary reliably with age.

3.2.4 Prior Media Experience

We next investigated whether children's prior media exposure (prior AR experience, general screen time, and media content type) predicted differences in their active exploration of AR. Due to the substantial skewness and floor effects in many variables, we applied data transformations as needed by dichotomizing highly skewed variables and creating composite indices, while retaining more evenly distributed variables in continuous form.

In particular, all six core AR experience measures (all binary and duration items) exhibited strong floor effects and right-skewed distributions. Social AR, Learning AR, and Game AR use was sparse, with over half of the children (51.7%, 49.0% and 60.3%, respectively) reporting no prior use; duration-based items showed even lower engagement levels (62.9%, 50.3%, and 60.9% at the lowest level, respectively). Given the limited variability, each binary item was dichotomized (No Experience = 0, Any Experience = 1) for analysis.

Measures of general screen time across five device categories displayed mixed distribution patterns. Television viewing was the most balanced, with only 5.3% reporting minimal use and 32.5% moderate levels of use. Tablet and phone use also showed variability across the scale. In contrast, Video Game consoles (66.2% minimal use), Computers (46.4%), and Other screens (62.3%) showed substantial floor effects. These skewed variables were dichotomized (No Experience = 0, Any Experience = 1), while Television viewing was analyzed in its original form.

Measures of media type exposure (animated versus live-action content) revealed that 23.8% of children were reported to consume almost all animated content, with 63.5% falling within the lowest two categories on the scale, indicating a general preference for animated media. Exposure to mixed-content media (e.g., animation blended with live-action) showed 19.9% reporting minimal exposure, with responses clustering at moderate frequency levels. As both variables reflected content type rather than usage intensity and exhibited relatively limited floor effects, they were retained in their original form for analysis.

Correlation analyses. Correlation analyses revealed no significant associations between children's prior media exposure (including television use, tablet/phone use, computer use, video game use, other screen use, AR-related experience, and media content type) and changes in active exploration from the pretest to the conflicting condition (all ps > .21). All correlations were small in magnitude ($|r_s| \leq .14$), indicating that children's media exposure patterns were not meaningful predictors of their tendency to adapt exploration strategies when encountering conflicting information in AR.

3.3 Discussion

Experiment 2 provides clear evidence that children's active exploration of AR is not driven by the mere appearance of AR content, but specifically by perceptual conflict. When children viewed perceptually consistent AR displays in the Pretest phase, they showed low levels of exploration and often accepted the AR-rendered image as corresponding to physical reality. This suggests that, in the absence of conflicting cues, children readily treat realistic AR content as veridical, highlighting their susceptibility to perceptual illusions created by synchronized overlays.

In contrast, when children encountered perceptual conflict, they engaged in more active and systematic exploration. Consistent with Experiment 1, this increase in exploration was accompanied by improved accuracy in identifying the real object. Moreover, children who actively explored were more likely to correctly identify the authentic image, reinforcing the idea that exploration serves as a mechanism for resolving perceptual ambiguity rather than reflecting mere engagement or attention.

Critically, however, evidence for lasting change was mixed. Although exploration in the Posttest condition did not differ significantly from the Conflicting condition, it also did not differ from the Pretest baseline, indicating that any carryover effects were limited or not robust. This suggests that perceptual conflict robustly triggers immediate epistemic engagement, but may not, on its own, produce stable or generalized changes in children's understanding of AR content over short timescales.

Age did not reliably predict overall exploration or accuracy, consistent with Experiment 1 and suggesting that the ability to use exploration to resolve perceptual ambiguity emerges early when task structure supports active inquiry. This pattern points to the importance of experiential and contextual factors, rather than purely maturational change, in shaping children's recognition of AR content. It further suggests that early-emerging reality-monitoring abilities may be latent but context-sensitive, becoming evident when children are encouraged to actively evaluate competing representations.

At the same time, age-related differences were observed in the strategies children used to explore. Younger children were more likely to rely on physical inspection of the object, whereas older children showed a relative shift toward causal testing of the AR system (i.e., inducing display "glitches"). This pattern suggests a developmental transition from direct, object-based

verification to more abstract reasoning about the generative mechanisms underlying digital representations. Given the lack of association between prior AR experience and behavior, this shift is unlikely to reflect familiarity with AR specifically, instead suggesting a more general developmental change in how children interrogate the relationship between representations and reality in novel technological contexts.

Finally, substantial individual variability in exploration and learning was not explained by prior AR or media experience, suggesting that responsiveness to perceptual conflict may depend on domain-general factors that differ across individuals such as curiosity or sensitivity to uncertainty. Together, these findings reinforce the role of perceptual conflict as a key driver of active information-seeking, while highlighting limits on the persistence of its effects.

4. General Discussion

AR integrates perceptually rich digital content with the physical environment, creating experiences that are vivid yet conceptually ambiguous. For young users, this presents both a challenge and an opportunity: AR can blur the boundary between appearance and reality, while also inviting inquiry into how digital information relates to the physical world. Understanding when children accept AR content at face value and when they interrogate its relation to reality is critical for theories of cognitive and symbolic development, as well as for the design of child-appropriate AR technologies.

Across two experiments, we found converging evidence that active exploration was driven not by the presence of AR digital overlays per se, but by perceptual conflict that disrupted expectations about what was real. Rather than merely capturing attention, perceptual conflict appeared to function as a cognitive “spotlight,” making discrepancies epistemically relevant. When appearances diverged across views, children did not passively observe the display; instead,

they adopted an investigative stance and generated actions, such as movement, manipulation, and cross-view comparison, that allowed them to whether the screen information reflected a physical object authentically or a representational transformation. Children who engaged in more active exploration were more likely to correctly identify what was physically present, highlighting exploration as a key mechanism linking perceptual conflict to accurate interpretation. This pattern aligns with theoretical accounts emphasizing the role of prediction error and disequilibrium in learning, in which violations of expectation destabilize existing assumptions and motivate reconstruction of understanding (Piaget, 1952), as well as with frameworks of conflict-monitoring, predictive processing, and error-driven causal learning that describe how discrepancies recruit control processes and interventions to reduce uncertainty (Botvinick et al., 2001; Friston, 2010; Gopnik & Schulz, 2004). Extending these accounts to AR contexts, the present findings show that purely representational discrepancies, even in the absence of physical anomalies, can elicit systematic exploration that supports accurate judgments about reality, a pattern not readily explained by confusion or novelty alone.

At the same time, the effects of perceptual conflict were better characterized as changes in how children engaged with information than as evidence of stable conceptual change. Although exploration increased from pretest to posttest, accuracy did not reliably improve, suggesting that brief encounters with conflict recalibrate inquiry strategies without fully reorganizing underlying concepts. This short-term epistemic recalibration is consistent with prior work showing that surprising outcomes can transiently increase the expected value of information, promoting exploration and hypothesis testing without producing durable conceptual change (Schulz & Bonawitz, 2007). These effects were observed across the sampled age range (3 to 6 years), indicating that the capacity to use exploration to resolve representational ambiguity is

present early in development. However, age-related differences emerged in how children explored: younger preschoolers more often relied on direct physical inspection, whereas older preschoolers increasingly engaged in causal testing that probed the AR system itself, such as inducing display "glitches." These differences were not explained by prior AR experience or general media exposure, suggesting that responsiveness to perceptual conflict may reflect broader cognitive or motivational factors, such as executive function, curiosity, or tolerance for uncertainty (Best & Miller, 2010; Diamond, 2013). Together, these findings suggest that variation in exploration reflects deeper differences in how perceptual conflict is interpreted as evidence about representation versus reality, motivating closer examination of perceptual conflict as an epistemic cue in symbolic contexts.

Before diving into the theoretical and applied implications of the findings, it is important to highlight the methodological contribution of the present work. The dual-screen AR diagram demonstrates how AR can serve not only as a stimulus to be evaluated, but as a tool for probing underlying cognitive mechanisms. By introducing controlled perceptual conflict without explicit instruction, the paradigm makes it possible to observe how children evaluate the relation between appearance and reality through embodied interaction in real time. This approach complements prior work that has primarily relied on verbal judgments or passive observation, and highlights the role of self-guided exploration in shaping symbolic understanding (Barsalou, 2020). More broadly, the dual-screen paradigm provides a framework for comparing how children interpret representations across media that vary in how they map onto the physical world, offering a pathway to investigate how symbolic interpretation generalizes across technologies that blur appearance and reality, with implications for the design of child-facing AR systems.

4.1 Perceptual conflict as an epistemic cue in AR contexts

How do perceptual conflicts in AR elicit children's exploratory behavior? From a symbolic-development perspective, perceptual conflict does not merely signal uncertainty about perceptual input; it destabilizes assumptions about the relation between appearance and reality. When perceptual input, authentic or artificial, aligned seamlessly with the physical world, children tended to accept the content at face value. In contrast, perceptual mismatch foregrounds the possibility that what is seen may not directly reflect physical reality, prompting children to test whether AR content behaves like a physical counterpart or like a representation governed by different rules. This pattern aligns with constructivist accounts of symbolic development, which emphasize that understanding representations emerges through active coordination of perception, action, and inference, rather than through exposure alone (Deloache, 2004; Gopnik & Schulz, 2004).

Why do these effects extend beyond momentary task demands? After encountering conflicts, children showed increased exploration on subsequent trials even in the absence of perceptual mismatch, despite no reliable improvement in accuracy. Features of AR that previously did not elicit exploration now prompted children to actively investigate the display. This carryover pattern is best understood as short-term epistemic recalibration, in which recent experience temporarily shifts how children evaluate visual information. Rather than reflecting durable conceptual change, symbolic understanding appears to be dynamically assembled and context-sensitive, becoming more salient when experience highlights the distinction between appearance and reality. Such transient recalibration is consistent with prior work showing that brief encounters with surprising or ambiguous evidence can alter learners' inquiry strategies and epistemic expectations without fully reorganizing underlying concepts (Schulz & Bonawitz, 2007).

At the same time, children varied in how they responded to perceptual conflict. Some showed increases in exploration, whereas others showed little change, suggesting that perceptual conflict does not uniformly function as an effective epistemic cue. One possibility is that children vary in their ability to detect discrepancies and recognize it as epistemically relevant (Lapidow et al., 2022). In visually complex or immersive environments, cognitive load may further limit children's capacity to analyze discrepancies or interpret them as informative (Makransky & Petersen, 2021). When the source of a discrepancy is unclear, some children may rely on surface appearance or intuitive judgments, particularly given ongoing development in distinguishing fantasy from reality (Flavell, 1986b; Woolley & E Ghossainy, 2013). In such cases, perceptual conflict creates an opportunity for inquiry, but whether it leads to investigation depends on whether it is interpreted as epistemically meaningful.

4.2 Embodied exploration as a pathway to symbolic understanding

How do children enact inquiry into symbolic content through embodied exploration? When visual information conflicts with expectations, children generate actions that allow them to investigate the relation between appearance and reality. In this context, exploration is not a generic response to uncertainty, but has a specific function that supports understanding the nature of symbolic content: Testing whether what is seen should be treated as a direct and authentic depiction of a physical object or as a representation generated by a system.

Why do children coordinate sensorimotor experience in the AR context differently? Younger preschoolers tend to rely on direct physical inspection, such as taking the card off and checking physical reality, which provides immediate and reliable sensory feedback. From a symbolic-development view, such strategies reflect an effort to resolve appearance-reality ambiguity by grounding interpretation in tangible evidence. This pattern is consistent with

accounts of embodied cognition, which emphasize that early conceptual understanding is rooted in sensorimotor experience (Barsalou, 2008; Piaget, 1952), as well as with findings that young children often struggle to coordinate symbolic representations with their referents when perceptual similarity is high (DeLoache, 1991, 2000; Flavell, 1986b). In AR contexts, where visual appearances can diverge from physical objects' true features, physical inspection offers a reliable way to determine what is physically real, rather than indicating a failure of symbolic understanding.

As children progress through the preschool years, their strategies increasingly reflect an emerging ability to reason about representational processes. That is, development shapes not whether children engage in symbolic interpretation, but how embodied action is organized to resolve representational ambiguity. Older preschoolers were more likely to engage in causal testing, generating actions that probed the system producing the perceptual discrepancy rather than the object itself. By manipulating viewpoints or inducing changes across displays, children could test whether transformations originated from the physical object or from the system. This shift aligns with developmental evidence that older preschoolers become increasingly capable of reasoning about hidden causes and distinguishing surface appearance from underlying structure (Gopnik & Schulz, 2004), as well as coordinating multiple representations and tracking relations between symbols and their referents (DeLoache, 1991; McNeil & Uttal, 2009). These strategies are further supported by advances in executive function and cognitive flexibility, which enable children to move beyond immediate perceptual cues and coordinate multiple perspectives (Best & Miller, 2010; Diamond, 2013).

Overall, these findings underscore that understanding the symbolic nature of AR does not depend on perception or action alone, but on how children organize embodied inquiry to resolve

representational ambiguity. Exploration provides a bridge between perceptual input and symbolic interpretation, allowing children to evaluate competing possibilities through action. The brief interaction window and minimal scaffolding highlight how children's symbolic interpretation can shift rapidly through action, while raising questions about how these moment-to-moment adjustments consolidate over repeated encounters or across media with different degrees of perceptual grounding.

4.3 Recommendations for child-centered AR technologies

The present findings have direct implications for the design of child-centered AR and other immersive technologies. Prior work has shown that visually fluent and immersive interfaces can reduce critical evaluation, particularly in young users (Makransky & Petersen, 2021). The current results suggest an alternative design principle: Rather than eliminating ambiguity, systems may better support understanding by revealing uncertainty in ways that invite inquiry. This aligns with research showing that users develop more calibrated trust when systems make their limitations observable and encourage exploration, rather than presenting outputs as seamless or authoritative (Dzindolet et al., 2003; Lee & See, 2004). The results also support broader HCI research demonstrating that users form more accurate mental models of interactive systems through interaction and feedback rather than passive observation (Norman, 2013).

From a design perspective, these findings highlight the value of epistemic affordances, features that make the constructed nature and limits of system outputs discoverable through interaction. Rather than relying solely on explicit labels or instructions, systems can be designed to support assumption testing through action, for example, by allowing users to compare alternative outputs or viewpoints. Such features shift children's engagement from passive viewing toward active evaluation, supporting more robust interpretation of AR content.

Developmental differences in exploration strategies further suggest that effective design should accommodate multiple forms of inquiry. Younger children more often relied on direct physical interaction, whereas older children increasingly tested the representational system itself. Accordingly, interfaces should support both concrete manipulation and opportunities to examine how representations change, allowing children to evaluate content through complementary strategies. Designing for this range of approaches aligns with user-centered principles emphasizing adaptability to diverse cognitive abilities and levels of expertise (Rogers et al., 2023).

Beyond system design, these findings also have implications for how children are supported in their interactions with AR. Children's understanding depends not only on exposure, but on whether contexts encourage them to reflect on and evaluate digital information. When perceptual discrepancies arise, adults can support children's sense-making by prompting them to compare views, explore alternative perspectives, or check physical counterparts. Importantly, such scaffolding does not require explicit technical explanations of AR. Based on our findings, even brief opportunities to notice and investigate discrepancies may recalibrate children's expectations about the reliability of visual information. In this way, AR can be positioned as a representational system to be actively evaluated, supporting critical engagement that extend beyond a single technology.

In conclusion, the present study shows that perceptual conflict in AR can shift children from passive acceptance toward active evaluation. When confronted with conflicting representations, children increase exploration to test how appearance relates to physical reality, revealing exploration as a key mechanism for interpreting visually convincing but ambiguous environments. These effects are best understood as short-term epistemic recalibration, in which

strategies for interpreting visual information are updated in response to recent experience, rather than as stable conceptual change. By showing how action supports the interpretation of emerging technologies, this work highlights the importance of designing interactive systems that make their underlying structure open to inquiry. Future research should examine whether these shifts persist over longer timescales and generalize across different forms of mixed reality.

Acknowledgments

Chapter 2 is currently being prepared for submission for publication: Lin, C., & Schachner, A. (In Prep). Perceptual conflict in augmented reality promotes exploration and supports reality judgments in young users.

REFERENCES

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *13*(3), 219–235.

Bahrick, L. E., & Watson, J. S. (1985). Detection of intermodal proprioceptive–visual contingency as a potential basis of self-perception in infancy. *Developmental Psychology*, *21*(6), 963–973.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(1), 617–645.

Barsalou, L. W. (2020). Challenges and opportunities for grounding cognition. *Journal of Cognition*, *3*(1), 31.

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, *81*(6), 1641–1660.

Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences*, *6*(6), 237–242.

Bodén, M., Dekker, A., Viller, S., & Matthews, B. (2013, June 24). Augmenting play and learning in the primary classroom. *Proceedings of the 12th International Conference on Interaction Design and Children*. IDC '13: Interaction Design and Children 2013, New York New York USA. https://doi.org/10.1145/2485760.2485767

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*(4), 215–234.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.

Brugman, H., & Russel, A. (2004). Annotating Multi-media/Multi-modal Resources with ELAN. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. https://aclanthology.org/L04-1285/

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349.

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, *26*(2), 197–223.

DeLoache, J. S. (1991). Symbolic functioning in very young children: Understanding of pictures and models. *Child Development*, *62*(4), 736.

DeLoache, J. S. (2000). Dual representation and young children's use of scale models. *Child Development*, *71*(2), 329–338.

Deloache, J. S. (2004). Becoming symbol-minded. *Trends in Cognitive Sciences*, *8*(2), 66–70.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*(1), 135–168.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718.

Flavell, J. H. (1986a). The development of children's knowledge about the appearance–reality distinction. *The American Psychologist*, *41*(4), 418–425.

Flavell, J. H. (1986b). The development of children's knowledge about the appearance-reality distinction. *The American Psychologist*, *41*(4), 418–425.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138.

Gopnik, A. (2012). Scientific thinking in young children: theoretical advances, empirical research, and policy implications. *Science (New York, N.Y.)*, *337*(6102), 1623–1627.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(30), 7892–7899.
Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, *8*(8), 371–377.

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R Package for generalized estimating equations. *Journal of Statistical Software*, *15*(2), 1–11.

Hood, B., Cole-Davies, V., & Dias, M. (2003). Looking and search measures of object knowledge in preschool children. *Developmental Psychology*, *39*(1), 61–70.

Kriegel, E. R., Lazarevic, B., Feifer, D. S., Athanasian, C. E., Chow, N., Sklar, J. P., Asante, Y. O., Goldman, C. S., & Milanaik, R. L. (2023). Youth and augmented reality. In *Springer Handbooks* (pp. 709–741). Springer International Publishing.

Lane, J. D., & Harris, P. L. (2014). Confronting, representing, and believing counterintuitive concepts: Navigating the natural and the supernatural. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *9*(2), 144–160.

Lapidow, E., Killeen, I., & Walker, C. M. (2022). Learning to recognize uncertainty vs. recognizing uncertainty to learn: Confidence judgments and exploration decisions in preschoolers. *Developmental Science*, *25*(2), e13178.

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Legare, C. H. (2012). Exploring explanation: explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development*, *83*(1), 173–185.

Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*, *81*(3), 929–944.

Lewry, C., Curtis, K., Vasilyeva, N., Xu, F., & Griffiths, T. L. (2021). Intuitions about magic track the development of intuitive physics. *Cognition*, *214*(104762), 104762.

Lin, Y., Stavans, M., & Baillargeon, R. (2022). Infants' physical reasoning and the cognitive architecture that supports it. In *The Cambridge Handbook of Cognitive Development* (pp. 168–194). Cambridge University Press.

Makransky, G., & Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review*, *33*(3), 937–958.

McNeil, N. M., & Uttal, D. H. (2009). Rethinking the use of concrete materials in learning: Perspectives from development and education. *Child Development Perspectives*, *3*(3), 137–139.
Moll, H., & Meltzoff, A. N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development*, *82*(2), 661–673.

Norman, D. A. (2013). *The design of everyday things*. MIT Press.

Oranç, C., & Küntay, A. C. (2019). Learning from the real and the virtual worlds: Educational use of augmented reality in early childhood. *International Journal of Child-Computer Interaction*, *21*, 104–111.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *The Behavioral and Brain Sciences*, *24*(5), 939–973.

Perez, J., & Feigenson, L. (2022). Violations of expectation trigger infants to search for explanations. *Cognition*, *218*(104942), 104942.

Piaget, J. (1952). *The origins of intelligence in children* (M. Cook, trans.). W W Norton & Co.

Rochat, P., & Striano, T. (2000). Perceived self in infancy. *Infant Behavior & Development*, *23*(3-4), 513–530.

Rogers, Y., Sharp, H., & Preece, J. (2023). *Interaction design: Beyond human-computer interaction* (6th ed.). John Wiley & Sons.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045–1050.

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing Our Lives with Immersive Virtual Reality. In *Frontiers in Robotics and AI* (Vol. 3). https://doi.org/10.3389/frobt.2016.00074

Stahl, A. E., & Feigenson, L. (2015). Cognitive development. Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312.

Troseth, G. L., & DeLoache, J. S. (1998). The medium can obscure the message: young children's understanding of video. *Child Development*, *69*(4), 950–965.

Troseth, G. L., Saylor, M. M., & Archer, A. H. (2006). Young children's use of video as a source of socially relevant information. *Child Development*, *77*(3), 786–799.

Tullos, A., & Woolley, J. D. (2009). The Development of Children's Ability to Use Evidence to Infer Reality Status. In *Child Development* (Vol. 80, Issue 1, pp. 101–114). https://doi.org/10.1111/j.1467-8624.2008.01248.x

Van Schijndel, T. J. P., & Raijmakers, M. E. J. (2016). Parent explanation and preschoolers' exploratory behavior and learning in a shadow exhibition: Parent explanation and preschoolers' exploration. *Science Education*, *100*(1), 153–178.

Woolley, J. D., & E Ghossainy, M. (2013). Revisiting the fantasy–reality distinction: Children as naïve skeptics. *Child Development*, *84*(5), 1496–1510.

CHAPTER 3 FROM MOVEMENT TO MEANING: PERCEIVED MUSICAL ENGAGEMENT IN A ROBOT ENHANCES SOCIAL EVALUATION OF IT

Humans infer intentions, emotions, and social meanings from movement. Even minimal motion cues can elicit inferences about agency, goals, and social interaction, as demonstrated by observers attributing intentions and relationships to moving geometric shapes (Heider & Simmel, 1944). Such inferences are foundational to social cognition because they shape how people evaluate and respond to others, including whether they are perceived as warm, competent, trustworthy, or threatening. Yet movement rarely specifies its meaning unambiguously, making interpretation fundamental to social evaluation.

Robots present a particularly pronounced case for understanding how social meaning is inferred from movement. As robots increasingly resemble humans in facial appearance, bodily form, and behavioral expressiveness, observers tend to interpret them as intentional agents (Broadbent, 2017; Epley et al., 2007). Due to their mechanical nature, identical robot movements may invite fundamentally different interpretations. A rhythmically moving robot, for instance, may be interpreted as responding meaningfully to music, executing a preprogrammed routine, or behaving mechanically. In Dennett's (1989) terms, observers may shift between interpretive stances when evaluating robot behavior. A physical stance explains movement as a product of mechanical processes, whereas a design stance interprets behavior in terms of programmed functions or algorithms. Meanwhile, a robot's humanlike appearance and expressive behavior may invite an intentional stance, inviting observers to interpret actions as reflecting goals, experiences, or internal states. Humanlike robots, therefore, occupy an unusual position between mechanical artifact and social agent in which identical movements may plausibly support competing interpretations, with potentially important consequences for social evaluation.

Musical movement provides a uniquely informative test case for understanding this inferential process because it is both socially meaningful and inherently ambiguous. Across cultures, music-related behaviors such as rhythmic movement, synchronization, and dance are deeply embedded in human social life and commonly associated with emotional expression, coordination, and social engagement (Brown, 1999; Cirelli et al., 2014; Clayton et al., 2005; Savage et al., 2020). Observers perceive correspondences between music and body movement, often treating rhythmic movement as signaling responsiveness to or engagement with music rather than arbitrary bodily motion (Davidson, 1993; Juchniewicz, 2008; Leman, 2019). Yet engagement with music is rarely directly observable and instead inferred from patterns of movement, synchrony, and contextual cues. The same rhythmic movement may be interpreted as genuine responsiveness to music, programmed synchronization, emotional expression, repetitive mechanics, or coincidence. Such ambiguity is especially consequential for humanlike robots, whose appearance invites intentional interpretation while their internal capacities remain fundamentally uncertain.

Understanding how observers interpret musically ambiguous robot movement is increasingly important as advances in robotics enable robots to exhibit behaviors suggestive of musical engagement, and is consequential for robots designed for companionship, caregiving, or entertainment, contexts in which responsiveness to music often carries social significance (Breazeal, 2002; Cirelli et al., 2014). Robots have been developed to perform musical instruments, improvise with human musicians, and move rhythmically in response to music (Hoffman & Weinberg, 2011; Petersen et al., 2010; Weinberg et al., 2020). Existing work, however, has focused primarily on engineering challenges such as movement generation, musical performance, and human-robot coordination. In parallel, research on human-robot

interaction has emphasized how morphology, facial realism, and anthropomorphic cues shape social evaluation (Carpinella et al., 2017; Eyssel & Hegel, 2012; Stroessner & Benitez, 2019). Comparatively little attention has been devoted to how observers interpret ambiguous robot behavior and whether such interpretations shape social evaluation. In particular, whether observers infer socially meaningful internal capacities from apparent engagement with music, and whether such inferences alter evaluations of humanlike robots, remains largely unknown.

We hypothesized that increasing levels of apparent musical engagement would enhance observers' social evaluations of a humanoid robot. Humanlike appearance and behavior encourage anthropomorphism, increasing the tendency to interpret robots in terms of human intentions, emotions, and internal states (Epley et al., 2007). Behaviors that appear socially meaningful would be likely to invite intentional interpretations, even when the underlying cause of the behavior remains uncertain. For instance, responsiveness to music is typically inferred rather than directly observed, based on movement patterns and contextual cues indicating engagement with music. If a humanoid robot appears to engage with music, whether through movement-music correspondence or contextual cues suggesting musical engagement, observers may infer emotional responsiveness, coordination, or socially meaningful internal capacities.

Observers readily infer expressive and emotional qualities from music-related movement. Even in the absence of sound, bodily movements associated with musical performance shape judgments of expressiveness, intentionality, and engagement (Davidson, 1993, 2001; Juchniewicz, 2008; Tsay, 2013). Moreover, coordinated and expressive movement communicates emotional states and interpersonal attunement (Atkinson et al., 2004). Accordingly, we expected social evaluations to vary systematically with the robot's apparent level of musical engagement: stronger evidence that the robot is meaningfully responding to

music should invite stronger inferences of emotional responsiveness, coordination, and socially meaningful internal capacities. Because perceptions of warmth and competence are shaped by inferred intentions and capabilities (Fiske et al., 2007), increasing apparent musical engagement was expected to predict greater perceived warmth and competence. Given that musical engagement provides a meaningful explanation for otherwise ambiguous movement, increasing apparent musical engagement was expected to reduce discomfort by making behavior appear less arbitrary or mechanical.

Alternatively, higher levels of apparent musical engagement may heighten negative reactions by amplifying a robot's human-like qualities. The uncanny valley hypothesis proposes that increasing human likeness does not always increase affinity: agents that appear highly but imperfectly human can instead evoke discomfort, eeriness, or aversion (Mori et al., 2012). One proposed mechanism is that conflicting perceptual and behavioral cues create uncertainty about agency, intentionality, or mental capacities (Kätsyri et al., 2015). For instance, when humanlike appearance is paired with movements or behaviors that violate perceptual expectations, observers may experience prediction errors or sensory conflict regarding how the agent should behave (Saygin et al., 2012; Urgen et al., 2018). Related accounts suggest that discomfort may arise when an entity occupies an ambiguous position between categorical boundaries, creating representational conflict about whether it should be understood as human or machine (Ferrey et al., 2015). Accordingly, stronger cues to musical engagement may intensify ambiguity by making the robot appear increasingly expressive and socially responsive while leaving its underlying mental capacities uncertain. Rather than rendering movement more interpretable, increasing apparent musical engagement may heighten uncertainty about whether the robot is genuinely responding to music or merely simulating human behavior. Therefore, higher apparent

musical engagement may paradoxically increase discomfort despite promoting more human-like interpretations.

The present research examined how cues to musical engagement shape social evaluations of a highly humanlike android. We manipulated apparent engagement with music while holding the android's physical movements constant, thereby isolating whether social evaluations depend on observable behavior itself or on how observers interpret that behavior (see Figure 3.1).

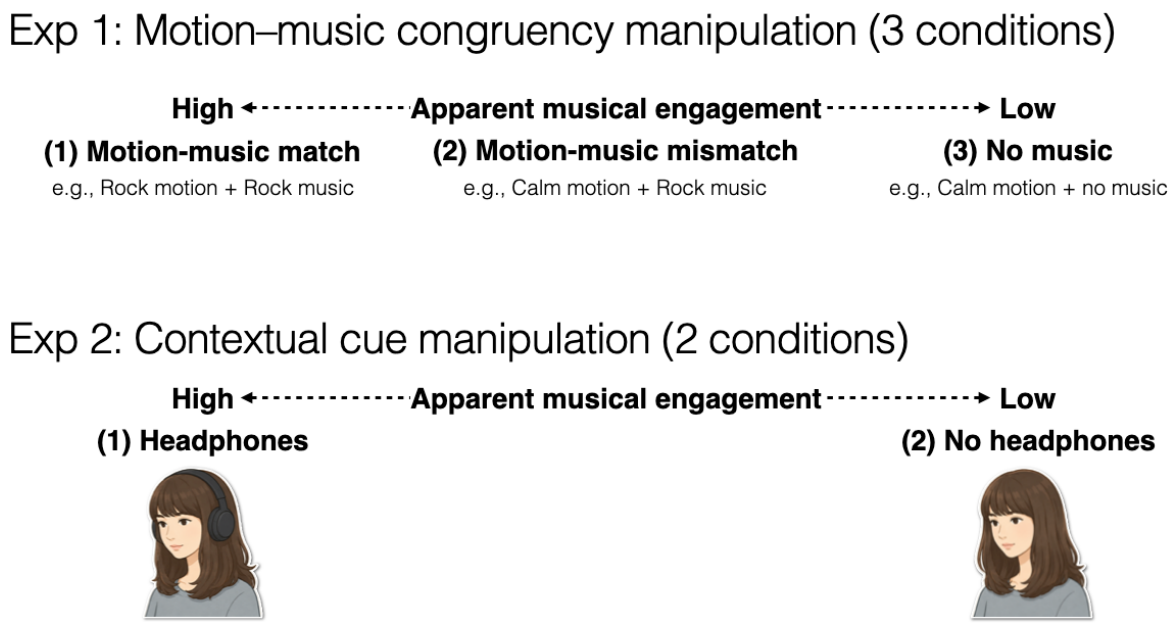

Figure 3.1: Experimental paradigms for Experiments 1 and 2. Experiment 1 manipulated the android's apparent musicality through music–motion congruency. Participants viewed identical robot movement profiles paired with matching music, mismatched music, or no music. Experiment 2 manipulated apparent musicality through contextual symbolic cues. Participants viewed the same movement profiles without accompanying audio, while the android either wore headphones or not, providing a contextual cue that the movements reflected musical engagement. Across both experiments, the robot's movements were held constant across conditions. Illustrations are schematic and depict only example stimuli.

In Experiment 1, participants viewed identical android movements that were synchronized with music, mismatched with music, or presented without music. We predicted that if observers infer musical engagement from movement-music correspondence, synchronized actions should increase perceived warmth and competence and reduce discomfort relative to

mismatched or absent music. Experiment 2 tested whether contextual cues alone could similarly shape interpretation and social attribution. Participants viewed the same movements without audible music, but the android either wore headphones or did not, providing a cue that the movement might reflect engagement with music. We predicted that if social evaluation depends not only on movement itself but also on inferred meaning, cues suggesting musical engagement should alter perceptions of warmth, competence, and discomfort even when music is absent. More broadly, these experiments investigate whether social evaluations of humanlike robots depend not only on observable behavior, but also on how observers interpret the underlying causes and capacities implied by that behavior.

Results

Across two experiments, we manipulated the android's apparent musicality through either behavioral coordination cues (music-motion matched, mismatched, or no music; Experiment 1) or contextual symbolic cues (headphone vs. no-headphone; Experiment 2) and assessed participants' social evaluations using the three subscales of the RoSAS: warmth, competence, and discomfort.

Experiment 1

As a manipulation check, we first examined whether participants spontaneously interpreted the android's behavior as music-related in open-ended descriptions. Responses were quantified using sentence embeddings projected onto a music-related semantic dimension, with higher scores indicating more music-related interpretations. Condition strongly predicted semantic musicality, $F(2, 189) = 77.07$, $p < .001$, $R^2 = .45$. Relative to the matched condition, participants in the mismatched condition described the android's behavior in less music-related terms, $b = -0.09$, $SE = 0.03$, $t(189) = -3.43$, $p < .001$, 95% CI [−0.14, −0.04], and participants in

the no-music condition showed substantially weaker music-related interpretations, $b = -0.31$, $SE = 0.03$, $t(189) = -12.05$, $p < .001$, 95% CI [−0.36, −0.26]. Participants most strongly interpreted the android's behavior as music-related in the matched condition, followed by the mismatched condition, with the weakest music-related interpretations in the no-music condition. See Figure 3.2.

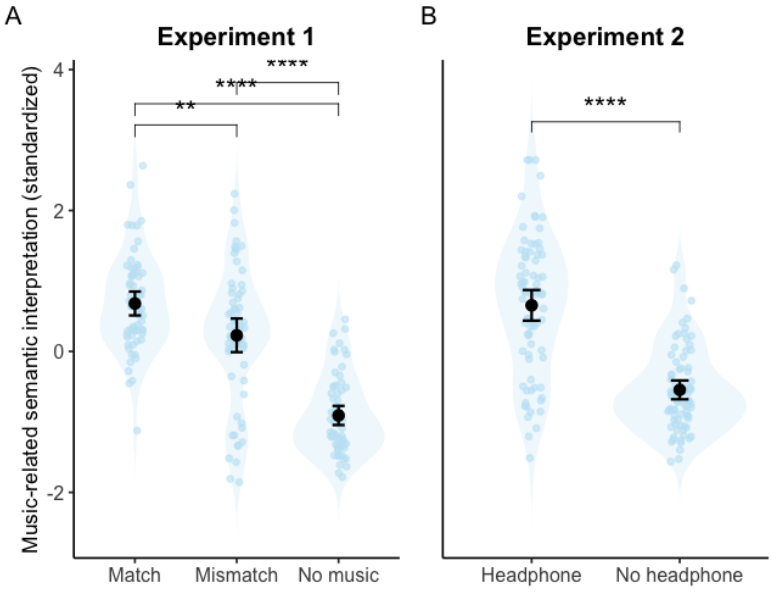

Figure 3.2: Music-related semantic interpretation across conditions in Experiments 1 and 2. Open-ended descriptions of the android's behavior were quantified using a standardized semantic musicality score. Participants in the matched condition described the android's behavior in significantly more music-related terms than participants in the mismatched and no-music conditions. Violin plots show score distributions; dots represent individual participants; points and error bars indicate means and 95% confidence intervals.

Does music-movement congruency shape social evaluations of the android?

Perceptions of the android varied systematically across musicality conditions. Welch one-way ANOVAs revealed significant effects of condition on perceived warmth, competence, and discomfort. Participants rated the android as warmer in the Match condition ($M = 4.45$, $SD = 1.78$) than in the Mismatch ($M = 3.49$, $SD = 1.60$; Games–Howell-adjusted $p = .005$) and No music conditions ($M = 3.36$, $SD = 1.63$; adjusted $p < .001$), whereas the latter two conditions did

not differ, adjusted p = .90. The omnibus effect was significant, Welch's $F(2, 125.7) = 7.46$, p = .003. Perceived competence was higher in the Match condition (M = 4.98, SD = 1.63) than in the Mismatch (M = 4.18, SD = 1.68; adjusted p = .018) and No-music conditions (M = 4.05, SD = 1.46; adjusted p = .002), with no difference between the latter two, p = .89, Welch's $F(2, 125.5) = 6.52$, adjusted p = .004. In contrast, discomfort showed the opposite pattern: participants reported lower discomfort in the Match condition (M = 3.87, SD = 1.57) than in the Mismatch (M = 4.56, SD = 1.68; adjusted p = .046) and No music conditions (M = 4.59, SD = 1.53; adjusted p = .026), which again did not differ (p = .99), Welch's $F(2, 125.8) = 4.26$, adjusted p = .016. See Figure 3.3.

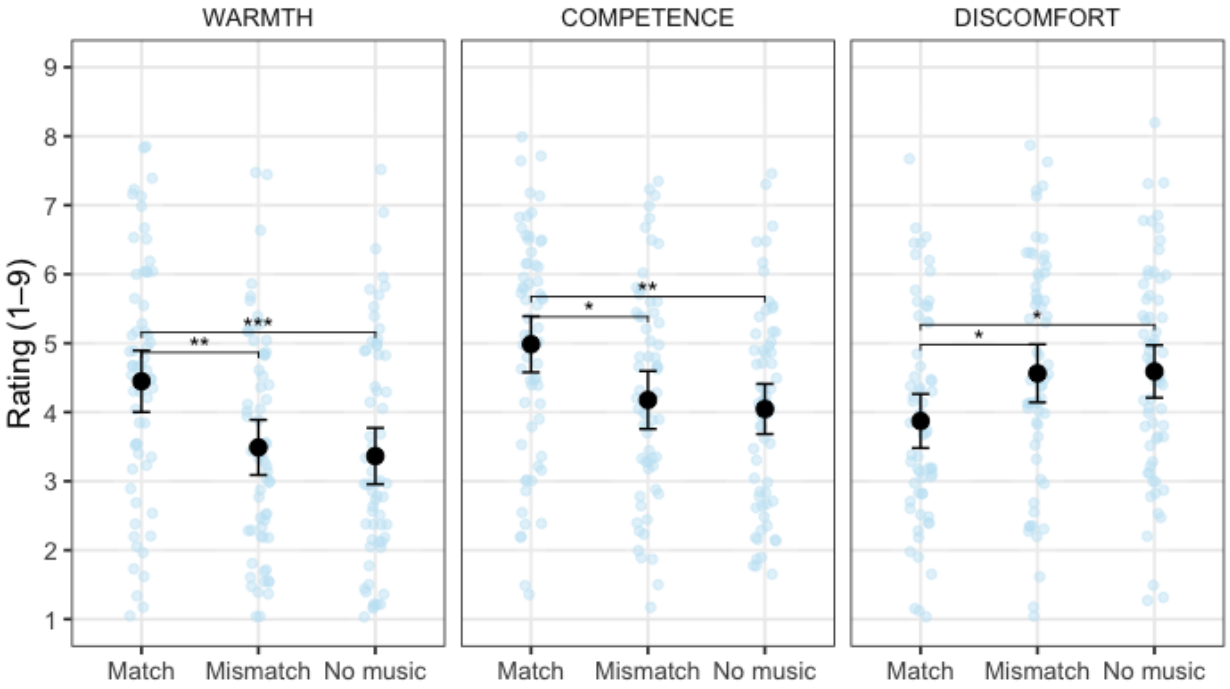

Figure 3.3: Social evaluations of the android as a function of musicality cues. Dots represent individual participants; points and error bars indicate means ± 95% confidence intervals. Significance brackets denote Games–Howell post hoc comparisons ($P < 0.05$). Matching movement-to-music cues increased perceived warmth and competence and reduced discomfort relative to the other conditions.

Are musically sophisticated participants more sensitive to robot musicality?

Moderation analyses examined whether musical sophistication altered sensitivity to apparent musicality cues. Adding musical sophistication alone did not improve model fit beyond experimental conditions for any outcome (all *p*s > .11). However, for competence, adding the interaction between condition and musical sophistication significantly improved model fit, $F(2, 186) = 4.29$, Holm-adjusted $p = .045$, and yielded the best-fitting model ($\Delta AIC = 2.72$ relative to the condition-only model). Participants lower in musical sophistication showed larger competence differences across conditions, whereas evaluations among more musically sophisticated participants were less sensitive to apparent musicality cues. Comparable moderation effects were not observed for warmth or discomfort (Holm-adjusted *p*s > 0.14; $\Delta AICs < 1$. See Figure 3.4.

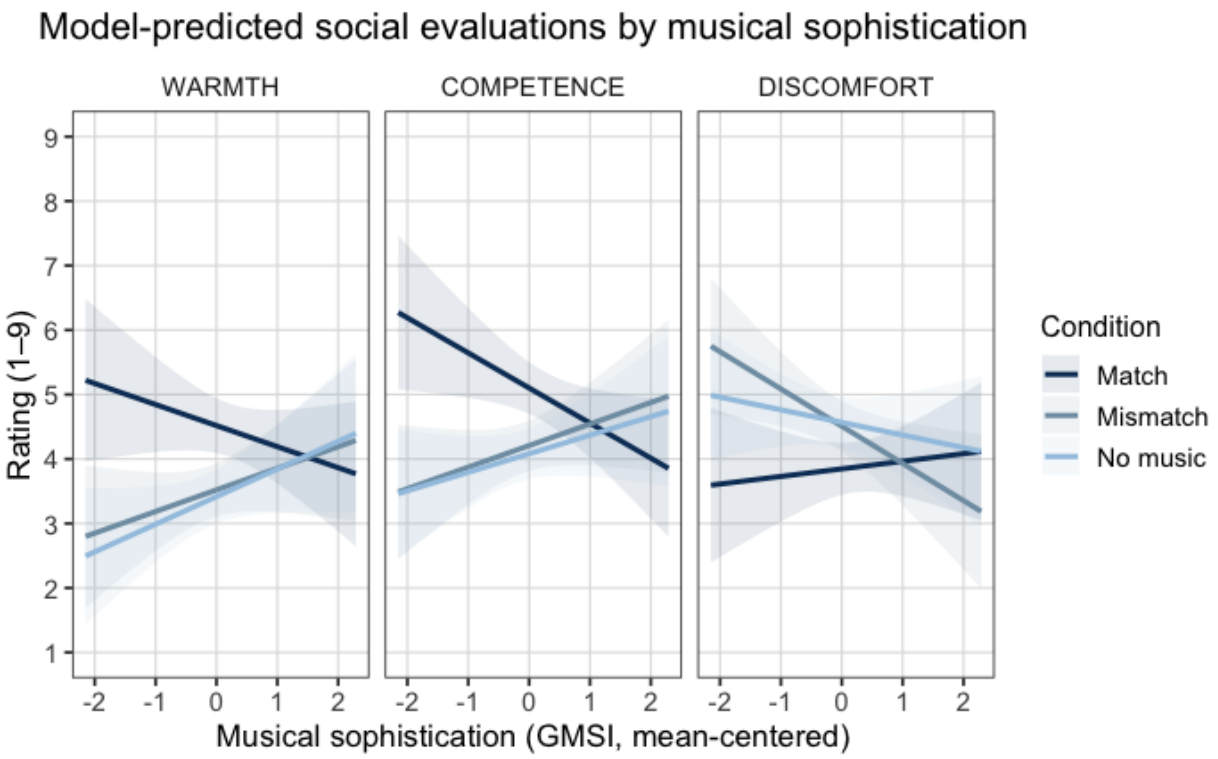

Figure 3.4: Model-predicted social evaluations as a function of musical sophistication in Experiment 1. Predicted ratings of warmth, competence, and discomfort are shown across levels of musical sophistication (GMSI, mean-centered) for the Match, Mismatch, and No music conditions. Lines and shaded ribbons indicate predictions and 95% confidence intervals from the interaction model (*Outcome* ~ Condition × GMSI), which was evaluated against simpler nested models. Musical sophistication moderated competence judgments but not warmth or discomfort after correction for multiple comparisons.

Are social evaluations mediated by inferred musical understanding?

We next examined whether social advantages of music–motion congruency were explained by participants' beliefs that the android could perceive and understand music ("inferred musical understanding"). Relative to the mismatched condition, participants in the matched condition attributed greater musical understanding to the android, $b = 0.52$, 95% CI [0.14, 0.87], $p = .005$. Inferred musicality positively predicted perceived warmth and competence (both ps $< .001$) but was not significantly associated with discomfort ($p = .079$). Mediation analyses revealed significant indirect effects of condition through inferred musicality for warmth

and competence, whereas evidence for discomfort was weaker and nonsignificant. Warmth retained a significant direct effect of condition after accounting for inferred musicality, indicating partial mediation, whereas the direct effect for competence was reduced to nonsignificance, suggesting that competence judgments were largely explained by inferences about the android's musical understanding. In contrast, reduced discomfort in the matched condition appeared to depend less on inferred musicality and more on other features of coordinated behavior. See Figure 3.5.

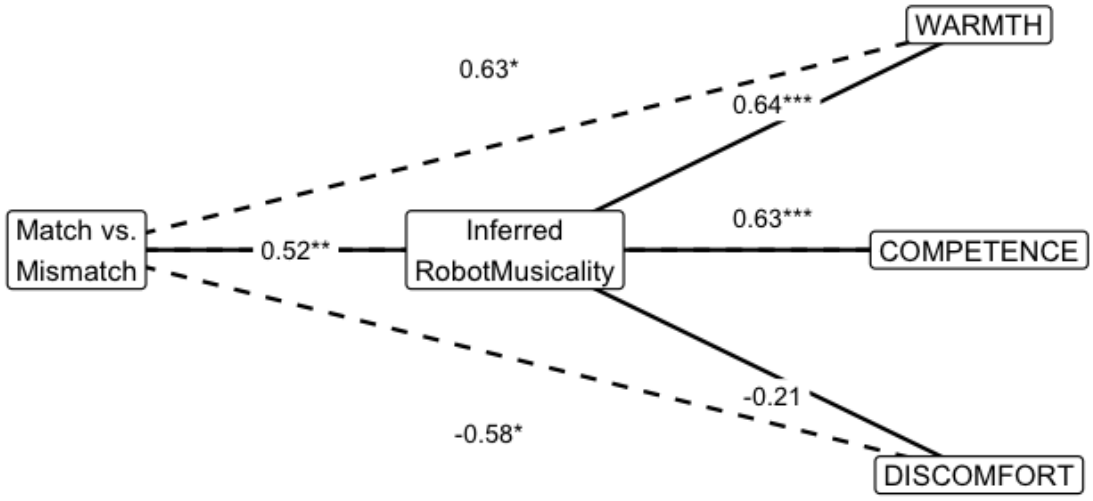

Figure 3.5: Mediation model of music–motion congruency effects on social evaluation in Experiment 1. Structural equation model testing whether effects of music–motion congruency (matched vs. mismatched) on perceived warmth, competence, and discomfort were mediated by participants' beliefs that the android could perceive and understand music ("inferred musical understanding"). Values indicate unstandardized path coefficients; dashed lines denote direct effects after accounting for the mediator ($P < 0.05$, $\mathbf{P} < 0.01$, $\boldsymbol{P} < 0.001$).

Experiment 2

Experiment 1 demonstrated that social evaluations of the android varied as a function of music–motion congruency, suggesting that observers inferred musical engagement from coordinated behavior. However, an important question remained: *Does social evaluation depend*

on dynamic behavioral evidence itself, or can minimal contextual cues similarly shape inferences about musical engagement? In Experiment 2, we tested whether a symbolic cue to musicality, the presence of headphones, would influence interpretations and social evaluations of the android in the absence of audible music. If observers use contextual information to infer meaningful engagement with music, then even minimal cues implying musicality should increase perceived warmth and competence while reducing discomfort.

Participants in the headphone condition described the android's behavior in significantly more music-related terms than participants in the no-headphone condition, $b = 0.24$, $SE = 0.03$, $t(158) = 9.36$, $p < .001$, 95% CI [0.19, 0.29], $R^2 = .36$. Thus, even in the absence of audible music, symbolic cues shifted spontaneous interpretations of the android's behavior toward musical engagement. See Figure. 3.2.

Does implied musical engagement shape social evaluations when audio is absent?

Social evaluations again varied as a function of musical context. Participants in the Headphone condition perceived the android as warmer ($M = 3.91$, $SD = 1.48$) than participants in the No headphone condition ($M = 3.25$, $SD = 1.47$), Welch's $t(158.2) = 2.82$, Holm-adjusted $p = .011$. Similarly, perceived competence was higher in the Headphone condition ($M = 4.19$, $SD = 1.43$) than in the No headphone condition ($M = 3.67$, $SD = 1.55$), Welch's $t(157.9) = 2.20$, Holm-adjusted $p = .029$. In contrast, discomfort was lower in the Headphone condition ($M = 4.24$, $SD = 1.71$) than in the No headphone condition ($M = 5.18$, $SD = 1.55$), Welch's $t(157.4) = -3.64$, Holm-adjusted $p = .001$. Together, these findings indicate that contextual cues (i.e., headphones) implying musical engagement, were associated with higher perceived warmth and competence while reducing discomfort toward the android. See Figure 3.6.

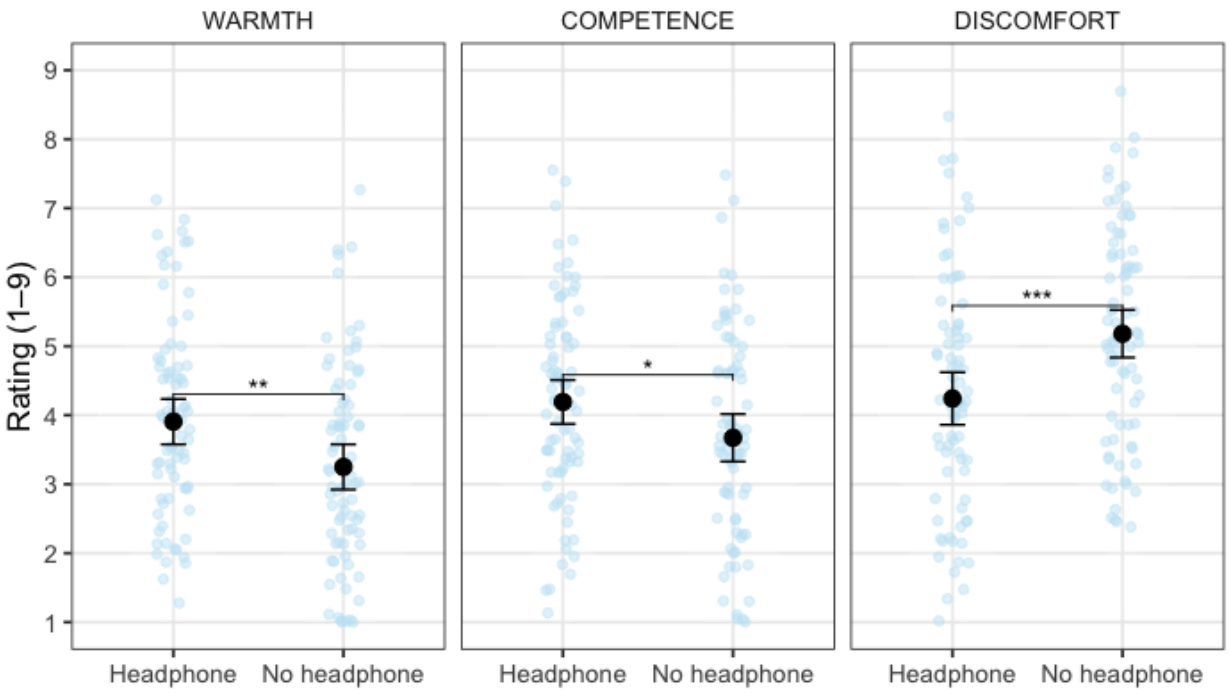

Figure 3.6: Social evaluations of the android across musical context conditions in Experiment 2. Dots represent individual participants, and points with error bars indicate means ± 95% confidence intervals. Significance brackets denote Welch independent-samples *t* tests ($P < 0.05$). Contextual cues supporting musical engagement increased perceived warmth and competence and reduced discomfort toward the android.

Are musically sophisticated participants more sensitive to robot musicality?

Moderation analyses examined whether musical sophistication altered sensitivity to apparent musicality cues. Adding musical sophistication alone did not improve model fit beyond experimental condition for warmth, competence, or discomfort (all *p*s > .38). Adding the interaction between condition and musical sophistication did not significantly improve prediction for warmth, competence, or discomfort (all *p*s > .18). Consistent with these inferential tests, condition-only models provided equivalent or better fit (ΔAICs < 2 for all comparisons). These findings suggest that social evaluations were not meaningfully moderated by participants' musical sophistication.

Are social evaluations mediated by inferred musical understanding?

We next examined whether the social effects of the headphone manipulation were explained by participants' beliefs that the android could perceive and understand music ("inferred musical understanding"). Participants in the headphone condition attributed greater musical understanding to the android than participants in the no-headphone condition, $b = 0.45$, 95% CI [0.13, 0.77], $p = .007$. Inferred musical understanding positively predicted perceived warmth, competence, and lower discomfort (all ps $< .05$). Mediation analyses revealed significant indirect effects of condition through inferred musical understanding for warmth and competence, whereas evidence for an indirect effect on discomfort did not reach significance. Warmth retained a significant direct effect of condition after accounting for inferred musical understanding, indicating partial mediation, whereas the direct effect for competence was reduced to nonsignificance, suggesting that competence judgments were largely explained by inferred musical understanding. In contrast, discomfort remained strongly associated with condition after accounting for the mediator, suggesting that reduced discomfort depended less on inferred musical understanding and more on other effects of contextual coherence. See Figure 3.7.

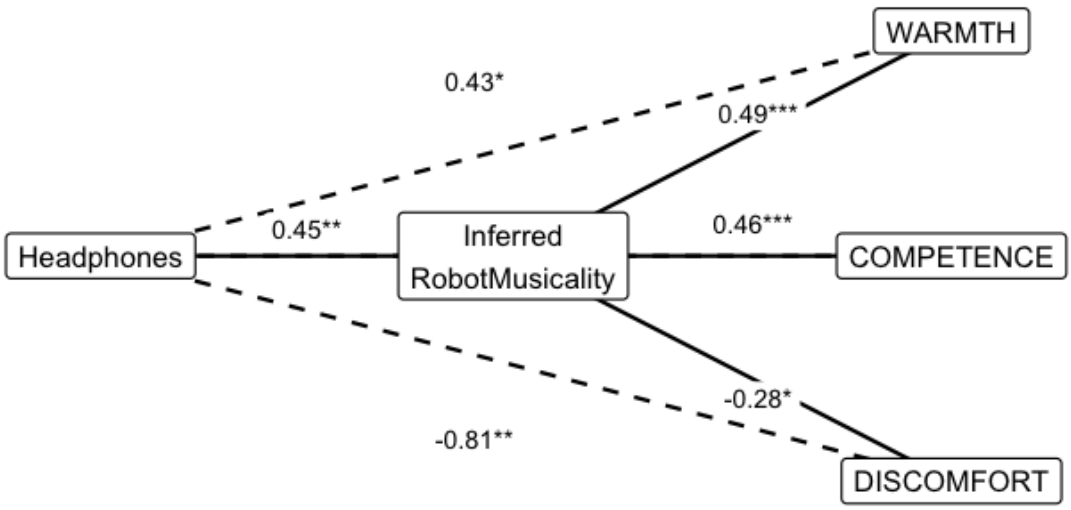

Figure 3.7: Mediation model of contextual musicality effects on social evaluation in Experiment 2. Structural equation model testing whether headphone cues influenced perceived warmth, competence, and discomfort through participants' beliefs that the android could perceive and understand music. Values indicate unstandardized path coefficients; dashed lines denote direct effects after accounting for the mediator ($P < 0.05$, $\mathbf{P} < 0.01$, $\boldsymbol{P} < 0.001$).

Discussion

The present study investigates how apparent musical engagement shapes social evaluations of a highly humanlike robot. Across two experiments, participants interpreted identical robot movements in increasingly music-related terms when cues to musical engagement were present. These effects were partly explained by participants' beliefs that the robot could perceive and understand music. When movement was synchronized with music, participants perceived the robot as warmer and more competent and experienced less discomfort than when the same movements were mismatched with music or presented without sound (Experiment 1). Similar effects emerged in the absence of audible music: Participants who viewed a robot wearing headphones, a contextual cue implying musical engagement, likewise evaluated the android more positively than those who viewed the same movements without headphones (Experiment 2). These findings suggest that social evaluations of highly humanlike robots were not solely driven by low-level perceptual properties of coordinated movement, such as

audiovisual fluency, rhythm, or perceptual synchrony, but also by observers' inferences about the underlying causes and internal capacities implied by that behavior.

These findings provide strong evidence that observers infer latent internal capacities from ambiguous behavior and use those inferences to guide social evaluation. Humans interpret behavior in relation to context, goals, and expectations to infer why an agent behaves as it does (Baker et al., 2009; Gergely & Csibra, 2003). Identical actions can acquire different meanings depending on contextual information, affecting judgments of intentionality, emotion, and agency (Barrett et al., 2011; Hassin et al., 2005). Humans infer musical engagement from movement, timing, contextual cues, and apparent coordination. Observers infer expressive intent even in the absence of sound, relying on bodily movement alone to judge musical engagement and responsiveness (Davidson, 1993; Tsay, 2013). Our findings extend this work by showing that such interpretive processes operate even when the actor is an artificial agent whose internal capacities remain uncertain. Such inferences may carry broader social significance because responsiveness to music is often associated with emotional sensitivity, attunement, and coordinated social behavior (Cirelli et al., 2014; Savage et al., 2020). A robot that appears capable of engaging with music may therefore be perceived not simply as moving rhythmically, but as exhibiting socially relevant forms of understanding or sensitivity.

The present findings also highlight an important but understudied feature of human-robot interaction: Humans evaluate not only what robots do, but what they believe robot behavior means. Humanlike robots occupy a unique social position in which observable behavior frequently exceeds certainty about underlying capacities. The same robot motion may be interpreted as emotionally expressive, mechanically scripted, socially responsive, or purely coincidental. In Dennett's terms (1989), observers may shift among physical, design, and

intentional stances when making sense of robot behavior. Experiment 1 alone could plausibly support such an account because synchronized music and movement may simply appear more coherent or pleasing than mismatched pairings. Experiment 2, however, indicates that interpretation itself plays a central role. In the absence of audible music, the mere presence of headphones shifted participants' spontaneous descriptions toward music-related interpretations and affected social evaluations. Because the robot's movements remained identical and no auditory information was available, these effects are difficult to explain solely through perceptual fluency or sensory congruency and suggest that even minimal symbolic cues can shape how observers explain ambiguous behavior, thereby altering social evaluation.

Musical sophistication showed limited evidence of moderating these effects. Although participants lower in musical sophistication appeared more sensitive to musicality cues when evaluating competence in Experiment 1, comparable moderation effects were absent for warmth and discomfort and did not replicate in Experiment 2. One possibility is that individuals with greater musical expertise hold more differentiated expectations regarding authentic musical responsiveness, making them less likely to treat surface-level synchronization as evidence of meaningful engagement (Carey et al., 2015; Danielsen et al., 2022). Alternatively, musically sophisticated individuals may attend to subtler mismatches between movement and music that nonexperts overlook (Senn et al., 2016). Future work using richer musical stimuli may clarify whether expertise meaningfully shapes social interpretations of robot behavior.

One intriguing aspect of the present findings concerns discomfort. Prior accounts of the uncanny valley propose that increasing human likeness may amplify discomfort when artificial agents evoke uncertainty regarding agency, intentionality, or internal states (Kätsyri et al., 2015; Mori et al., 2012). Humanlike robots can create ambiguity when appearance suggests social

capacities that behavior fails to support, producing prediction errors or representational conflict (Ferrey et al., 2015; Saygin et al., 2012; Urgen et al., 2018). Stronger cues to musical engagement, thereby, might plausibly have heightened discomfort by making the robot appear increasingly expressive while leaving its actual capacities unresolved. We observed the opposite pattern: Cues supporting musical engagement consistently reduced discomfort, even though the robot's underlying capacities remained fundamentally uncertain.

One explanation is that musical engagement supplied a coherent interpretive framework for otherwise ambiguous behavior. The same bodily movements may have appeared arbitrary or mechanical when disconnected from meaningful context, but became more intelligible when interpreted as responses to music. This interpretation aligns with predictive accounts of perception and social cognition, which propose that humans continuously generate explanations for others' behavior and experience uncertainty or discomfort when expectations remain unresolved (Friston & Kiebel, 2009; Van de Cruys et al., 2014). Apparent musical engagement may have reduced discomfort, not because participants became convinced that the robot genuinely possessed humanlike inner experiences, but because its behavior became easier to explain. In other words, discomfort toward highly humanlike robots may depend not simply on human likeness itself, but on whether observers can construct a sufficiently coherent account of what the robot is doing and why.

The present findings further suggest that positive social evaluation and discomfort may arise through partially distinct cognitive mechanisms. Whereas warmth and competence judgments were consistently mediated by beliefs about the robot's musical understanding, evidence for the mediation of discomfort was weaker and less consistent across experiments. Participants in musically congruent or headphone conditions still reported lower discomfort even

after accounting for inferred musicality. This pattern raises the possibility that reduced discomfort depends less on attributing sophisticated internal capacities and more on increased interpretability or behavioral coherence. That is, humans may not need to believe that a robot genuinely understands music in order to feel more comfortable with it. Discomfort may diminish once behavior appears sufficiently meaningful, predictable, or contextually appropriate. Such a distinction may help reconcile inconsistent findings in the uncanny valley literature by suggesting that positive evaluations and aversion need not emerge from identical inferential processes.

Several limitations should also be considered. First, the present experiments focused on a single highly humanlike robot, Erica. Highly anthropomorphic robots may invite stronger intentional interpretations than mechanical or abstract agents, potentially affecting how musical engagement is inferred. Second, participants evaluated prerecorded videos rather than engaging in live interactions. Real-time interaction may affect the inferential processes observed here. Third, the current studies focused specifically on musical engagement, a domain strongly associated with emotional expression, coordination, and social bonding. Whether similar effects emerge for other socially meaningful behaviors remains an open question. Future work should investigate whether comparable effects emerge in other motion types involving ambiguous responsiveness, such as gaze coordination or responsiveness to social touch, and whether similar effects emerge for less anthropomorphic robots.

The present findings carry broad implications for the design of socially acceptable robots. Humanlike robots increasingly appear in caregiving, educational, therapeutic, and companionship contexts in which behavior can be socially ambiguous and internal capacities are difficult to assess. Existing approaches to robot design often emphasize improving movement

realism, facial expressiveness, or behavioral sophistication. Our findings suggest that social acceptance may depend not only on what robots do, but also on how their actions are interpreted. Small contextual or behavioral cues can meaningfully alter whether robot behavior appears coherent, socially responsive, or unsettling. Social robot designers may therefore benefit from considering the interpretive frameworks their systems invite. A robot that transparently communicates why it behaves in certain ways, or provides contextual cues that render its actions understandable, may evoke more positive responses even when underlying capabilities remain limited. Designing socially acceptable robots may therefore require attention not only to behavioral realism, but also to interpretability.

In conclusion, the present work demonstrates that social evaluations of highly humanlike robots depend not only on observable behavior itself but also on how observers interpret that behavior. Across two experiments, identical robot movements were evaluated differently depending on whether participants construed them as meaningful responses to music. Behavioral congruency and minimal contextual cues systematically altered spontaneous interpretations, beliefs about the robot's musical understanding, and downstream social evaluations. These findings suggest that humans do not simply react to robot behavior, but infer what behavior implies about underlying capacities and responsiveness. As social robots become integrated into everyday social life, understanding how people interpret ambiguous behavior may prove as important as engineering the behavior itself.

Materials and Methods

Stimuli. Stimuli featured the humanoid android Erica, developed by the Japan Science and Technology Agency (JST) at Osaka University and the Advanced Telecommunications Research Institute International (ATR) at Kyoto University (Glas et al., 2016). Erica has been described as

one of the world's most humanlike androids, with silicone-resin skin and a facial appearance modeled from 30 images of Japanese and European women in their twenties (McCurry, 2015).

To create videos in which Erica's movements appeared to be realistic responses to music, we modeled her actions after videos of adult women YouTubers listening and reacting to music. These videos were selected for their expressive responses to music and for representing two distinct musical styles. One clip featured responses to the energetic rock song *Child in Time* by Deep Purple (1970), whereas the other featured responses to the calm piano piece *Visions* from the album *Journey Back to Sedona* by Barabas (1991).

Erica's movement profiles were programmed to mimic the temporal dynamics of the human models' responses on a frame-by-frame basis. The robot was then filmed against a light-colored curtain backdrop, with only the head and upper torso visible. Within each movement clip, Erica's body movements were held constant, with minor appearance variations limited to whether her eyes remained open or closed during movement and, in Experiment 2, whether the robot wore headphones.

Each final stimulus video consisted of two movement profiles corresponding to the two musical styles and lasted approximately 1 min in total. To reduce potential order and appearance effects, each condition included four stimulus variants that counterbalanced (i) the order of movement profiles (rock first vs. calm first) and (ii) whether the robot's eyes were open or closed during the movements. In Experiment 2, stimulus variants additionally varied whether Erica wore headphones. The order of movement profiles and accompanying soundtracks was edited in iMovie.

General Procedure and Design.

All experiments were approved by the university institutional review board (IRB protocol #161173). Participants were undergraduate students recruited through the psychology department participant pool (SONA system) and received course credit for participation. Each participant was recruited for only one of the experiments. The experiments were administered online using Qualtrics. Participants provided informed consent prior to participation. For each experiment, sample size, procedures, hypotheses, and analysis plans were preregistered before data collection.

In each experiment, participants were randomly assigned to an experimental condition in a between-subjects design. Procedures were otherwise identical across conditions. Participants first viewed a video clip, and a JavaScript script prevented them from advancing until the video finished playing.

Following video viewing, participants completed a free-response manipulation check in which they described what the android was doing in the video. Participants then completed the Robotic Social Attributes Scale (RoSAS; Carpinella et al., 2017) and a subscale of the Goldsmiths Musical Sophistication Index (Gold-MSI; Müllensiefen et al., 2014), as described in Measures. Items within each scale were presented in randomized order.

Participants then completed demographic questions, including age, gender, ethnicity, and educational background. To assess potential demand characteristics, participants responded to an open-ended question asking them to guess the study's purpose. To identify potentially invalid or automated responses, participants were asked to describe what they had done during the study in a free-response format. As an attention check, participants indicated whether they had heard the android speak in the video. Finally, as an additional manipulation check, participants rewatched

the video and rated the extent to which they believed the android could perceive and understand music.

Measures. Participants completed measures assessing social perception of the android, individual musical sophistication, and beliefs about the android's musicality.

Social perception of the android. Participants' social perception of the android was assessed using the Robotic Social Attributes Scale (RoSAS; Carpinella et al., 2017). The RoSAS consists of 18 items measuring three dimensions of robot perception: warmth (e.g., "compassionate"), competence (e.g., "interactive"), and discomfort (e.g., "scary"), with six items per subscale. Participants rated the extent to which each attribute described the android shown in the video on a 9-point Likert scale ranging from 1 (Definitely not associated) to 9 (Definitely associated). Internal reliability was high across subscales (Cronbach's $\alpha = .88$ for warmth, .86 for competence, and .80 for discomfort).

Participants' musical sophistication. Participants' musical sophistication was assessed using the General Musical Sophistication (GMS) scale from the Goldsmiths Musical Sophistication Index (Gold-MSI; Müllensiefen et al., 2014). The GMS scale comprises 18 items that provide an overall index of musical sophistication. Participants responded using a 7-point Likert scale. Internal reliability for the GMS score was high (Cronbach's $\alpha = .87$).

Belief in the android's musicality. To assess participants' beliefs about the android's musicality, participants responded to a single item: "To what extent do you think that the android can perceive and understand music?" Responses were recorded on a 5-point Likert scale ranging from 1 (None at all) to 5 (A great deal).

Experiment 1

Participants

In line with our preregistered sample size, 192 adults participated in Experiment 1, with 64 in each condition (age: M = 21.28 years, SD = 2.00, range = 19–38). The sample included 141 women (73.4%), 50 men (26.0%), and 1 participant identifying as another gender category (0.5%). Participants reported diverse racial backgrounds, including Asian (46.9%), White (29.2%), multiracial or other racial identities (18.7%), American Indian or Alaska Native (2.1%), and Black or African American (1.0%). Additionally, 33.9% of participants identified as Hispanic, Latino, or of Spanish origin. Educational attainment varied, with most participants reporting some college without a degree (44.8%) or a high school diploma or equivalent (37.5%), followed by an associate degree (9.4%) and a bachelor's degree (8.3%).

Conditions

We manipulated the android's apparent musicality through behavioral coordination cues across three conditions: (1) music-motion matched, in which the two musical clips corresponded to the robot's movement profiles; (2) music-motion mismatched, in which the musical clips were paired with noncorresponding movement profiles (i.e., the rock and calm soundtracks were swapped); and (3) no music, in which the robot exhibited the same movements without accompanying audio. The robot's movements were identical across conditions, such that only the auditory context varied.

Experiment 2

Participants

In line with our preregistered sample size, 160 adults participated in Experiment 2, with 80 participants in each condition (age: M = 21.29 years, SD = 1.94, range = 19–34). The sample included 121 women (75.6%), 38 men (23.8%), and 1 participant identifying as another gender category (0.6%). Participants reported diverse racial backgrounds, including Asian (50.6%),

White (27.5%), multiracial or other racial identities (16.9%), and Black or African American (3.8%). Additionally, 23.1% of participants identified as Hispanic, Latino, or of Spanish origin. Educational attainment varied, with most participants reporting some college education without a degree (41.2%) or a high school degree/equivalent (35.0%), followed by an associate degree (16.9%) and a bachelor's degree (6.9%).

Conditions

We manipulated the android's apparent musicality through contextual symbolic cues across two conditions: (1) headphone, in which the robot wore headphones while exhibiting music-responsive movements, and (2) no-headphone, in which the robot exhibited the same movements without headphones. To isolate the effects of contextual cues to musical engagement, all audio was removed from the stimulus videos. The robot's movements were identical across conditions, such that only the visual cue (i.e., the presence or absence of headphones) varied.

Acknowledgments

Chapter 3 is currently being prepared for submission for publication: Lin, C., & Schachner, A. (In Prep). From movement to meaning: perceived musical engagement in a robot enhances social evaluation of it.

REFERENCES

Atkinson, A. P., Dittrich, W. H., Gemmell, A. J., & Young, A. W. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, *33*(6), 717–746.

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

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, *20*(5), 286–290.

Breazeal, C. (2002). Designing Sociable Machines. In K. Dautenhahn, A. Bond, L. Cañamero, & B. Edmonds (Eds.), *Socially Intelligent Agents: Creating Relationships with Computers and Robots* (pp. 149–156). Springer US.

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology*, *68*(1), 627–652.

Brown, S. (1999). The "musilanguage" model of music evolution. In *The Origins of Music*. The MIT Press.

Carey, D., Rosen, S., Krishnan, S., Pearce, M. T., Shepherd, A., Aydelott, J., & Dick, F. (2015). Generality and specificity in the effects of musical expertise on perception and cognition. *Cognition*, *137*, 81–105.

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The Robotic Social Attributes Scale (RoSAS): Development and Validation. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 254–262.

Cirelli, L. K., Einarson, K. M., & Trainor, L. J. (2014). Interpersonal synchrony increases prosocial behavior in infants. *Developmental Science*, *17*(6), 1003–1011.

Clayton, M., Sager, R., & Will, U. (2005). In time with the music: the concept of entrainment and its significance for ethnomusicology. *European Meetings in Ethnomusicology.*, *11*, 1–82.

Danielsen, A., Nymoen, K., Langerød, M. T., Jacobsen, E., Johansson, M., & London, J. (2022). Sounds familiar(?): Expertise with specific musical genres modulates timing perception and micro-level synchronization to auditory stimuli. *Attention, Perception & Psychophysics*, *84*(2), 599–615.

Davidson, J. W. (1993). Visual Perception of Performance Manner in the Movements of Solo Musicians. *Psychology of Music*, *21*(2), 103–113.

Davidson, J. W. (2001). The role of the body in the production and perception of solo vocal performance: A case study of Annie Lennox. *Musicae Scientiae: The Journal of the European Society for the Cognitive Sciences of Music*, *5*(2), 235–256.

Dennett, D. C. (1989). *The Intentional Stance*. MIT Press.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886.

Eyssel, F., & Hegel, F. (2012). (S)he's got the look: Gender stereotyping of robots[1]. *Journal of Applied Social Psychology*, *42*(9), 2213–2230.

Ferrey, A. E., Burleigh, T. J., & Fenske, M. J. (2015). Stimulus-category competition, inhibition, and affective devaluation: a novel account of the uncanny valley. *Frontiers in Psychology*, *6*, 249.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1521), 1211–1221.

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

Glas, D. F., Minato, T., Ishi, C. T., Kawahara, T., & Ishiguro, H. (2016). ERICA: The ERATO Intelligent Conversational Android. *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 22–29.

Hassin, R. R., Aarts, H., & Ferguson, M. J. (2005). Automatic goal inferences. *Journal of Experimental Social Psychology*, *41*(2), 129–140.

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Hoffman, G., & Weinberg, G. (2011). Interactive improvisation with a robotic marimba player. *Autonomous Robots*, *31*(2), 133–153.

Juchniewicz, J. (2008). The influence of physical movement on the perception of musical performance. *Psychology of Music*, *36*(4), 417–427.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, *6*, 390.

Leman, M. (2019). *Embodied music cognition and mediation technology*. MIT Press. https://doi.org/10.7551/mitpress/7476.001.0001

McCurry, J. (2015, December 31). Erica, the “most beautiful and intelligent” android, leads Japan’s robot revolution. *The Guardian*.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine / IEEE Robotics & Automation Society*, *19*(2), 98–100.
Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). Measuring the facets of musicality: The Goldsmiths Musical Sophistication Index (Gold-MSI). *Personality and Individual Differences*, *60*, S35.

Petersen, K., Solis, J., & Takanishi, A. (2010). Musical-based interaction system for the Waseda Flutist Robot: Implementation of the visual tracking interaction module. *Autonomous Robots*, *28*(4), 471–488.

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2020). Music as a coevolved system for social bonding. *The Behavioral and Brain Sciences*, 1–42.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience*, *7*(4), 413–422.

Senn, O., Kilchenmann, L., von Georgi, R., & Bullerjahn, C. (2016). The effect of expert performance microtiming on listeners' experience of groove in swing or funk music. *Frontiers in Psychology*, *7*, 1487.

Stroessner, S. J., & Benitez, J. (2019). The social perception of humanoid and non-humanoid robots: Effects of gendered and machinelike features. *Advanced Robotics: The International Journal of the Robotics Society of Japan*, *11*(2), 305–315.

Tsay, C.-J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(36), 14580–14585.

Urgen, B. A., Kutas, M., & Saygin, A. P. (2018). Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia*, *114*, 181–185.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychological Review*, *121*(4), 649–675.

Weinberg, G., Bretan, M., Hoffman, G., & Driscoll, S. (2020). *Robotic Musicianship: Embodied Artificial Creativity and Mechatronic Musical Expression*. Springer, Cham.

DISCUSSION

How do humans use perceptual representations as evidence to infer underlying states in technology-mediated environments? Across the social, physical, and agentive domains of cognition, the present dissertation suggests that successful adaptation depends on determining what mediated information means and how it relates to reality. In video chat (Chapter 1), children inferred what another person could see despite disrupted spatial correspondences among the self, the screen, and the camera. In AR (Chapter 2), children evaluated whether perceptual features reflected properties of physical objects or digital augmentation. In the social evaluation of a humanlike robot (Chapter 3), adults assessed social qualities and internal capacities using behavioral cues that only indirectly reflected underlying states. These contexts share a cognitive challenge: determining how perceptual information maps onto reality and using that relationship to guide behavior, inference, and judgment.

I proposed that cognitive adaptation to such environments depends on representation-reality alignment (RRA), the process of evaluating how perceptual information corresponds to underlying states and using this relationship to guide behavior. Findings in Chapter 1 (video chat) and Chapter 2 (AR) reveal a developmental change that reflects increasing flexibility in how perceptual information is interpreted and used. Younger children often relied on immediately available perceptual cues, treating mediated environments as though ordinary perception-action relations remained intact. Older children, in contrast, increasingly treated perceptual information as evidence requiring interpretation, using action and prior experience to infer underlying states that were not directly observable. This inferential process may extend beyond childhood and across domains. In Chapter 3, adults appear to draw on prior social and embodied knowledge to interpret ambiguous signals in robot motion. Together, these findings

support the proposal that developmental change in RRA reflects increasing flexibility in treating perceptual representations as evidence about reality rather than reality itself. Across domains, successful performance depended on coordinating perception, inference, and, in some contexts, action to determine how mediated information should be interpreted and whether it could be trusted.

The following sections consider the mechanisms that may support this process and discuss the broader theoretical and applied implications of RRA for understanding cognition in increasingly mediated environments.

1. Interpreting and calibrating technology-mediated information

The present findings suggest that adaptation to mediated environments may depend on learning how to interpret and calibrate perceptual information when reality is partially obscured, transformed, or distributed across representational systems. In face-to-face interaction, perceptual information often supports tightly coupled perception-action loops. Visual access, object properties, and social behaviors typically correspond directly to underlying states, allowing perception to guide behavior with relatively little ambiguity. Technology-mediated environments alter these relations. Screens, cameras, digital overlays, and robotic systems introduce representational layers that can transform what is perceived and how perceptual information should be interpreted. Successful interactions, therefore, depend on determining how that information maps onto the relevant underlying reality.

The developmental findings from Chapter 1 (video chat) and Chapter 2 (AR) reveal a developmental shift in how children approach this challenge. Younger children often appeared to treat mediated information as though ordinary perceptual assumptions remained intact. In video chat, some children attempted to show objects toward the partner's on-screen face when self-

view was unavailable, suggesting reliance on immediately visible spatial cues rather than on the interface's representational logic. Older children became more likely to use self-view information to infer another person's visual access, suggesting greater sensitivity to how perceptual representations correspond to another person's perspective. In AR, younger children were more likely to rely on perceptual appearance in the real world when determining whether an object was physically real or digitally augmented. Older children increasingly departed from this strategy, as evidenced by engaging in exploratory behaviors that generated abstract causal effects of movements and perceptual information. Across both domains, developmental changes appeared to reflect increasing ability to treat perceptual information as evidence about hidden states rather than as direct access to reality. Importantly, this shift emerged even when children possessed little explicit knowledge of how the relevant technologies operated, suggesting that adaptation depended less on technical understanding than on increasingly flexible interpretation of mediated information.

The present findings further suggest that action plays an important epistemic role in this process. Across studies, children often acted to determine how mediated representations corresponded to reality. In video chat, scaffolding supported children's ability to coordinate actions with self-view information, helping reveal how movements in the camera frame related to another person's visual access. In AR, perceptual conflict increased exploratory behaviors such as object removal and movement across viewpoints, which appeared to generate information about the stability and source of perceptual features. These findings suggest that action can serve as a mechanism for constructing understanding.

From an RRA perspective, action supports the calibration through iterative cycles of perception, inference, and testing. When perceptual information is uncertain or conflicting,

observers may generate competing interpretations regarding underlying reality and use action to evaluate these possibilities. Successful adaptation, therefore, involves learning how to coordinate perception and action to determine when mediated information should be trusted, ignored, or further investigated. Across Chapters 1 and 2, action served as a direct means of calibrating representation–reality mappings. In contrast, Chapter 3 highlights a related process that relied more heavily on interpretive inference. Adults did not actively test competing explanations of robot behavior, but instead evaluated behavioral cues in relation to prior knowledge and expectations. Together, these findings suggest that RRA may be supported through different combinations of action and inference depending on the demands of a particular domain.

This interpretation is broadly compatible with inferential accounts of cognition, including predictive and active inference frameworks (Clark, 2019; Friston, 2010). These accounts propose that perception involves inferring hidden causes under uncertainty and that action can reduce uncertainty by sampling informative evidence. Although the present studies were not designed to test predictive mechanisms directly, the findings suggest that mediated environments may amplify inferential demands by weakening ordinary correspondences between perception and reality. In video chat, children's increasing ability to coordinate self-view with another person's perspective may reflect developmental improvements in reasoning about a partner's visual access. In AR, perceptual conflict prompted exploratory behavior that generated evidence relevant to the evaluation of competing interpretations. Across contexts, adaptation appeared to depend on using perceptual information as evidence rather than accepting it at face value.

One possible mechanism supporting this process is the recruitment of embodied and self-based knowledge to scaffold interpretation. When perceptual information becomes uncertain, observers may rely on their own perceptual, motor, and social experiences to infer how mediated

representations correspond to underlying states. Evidence for this possibility was strongest in video-mediated interaction. Children's increasing ability to use self-view to infer a partner's visual access is broadly consistent with developmental accounts proposing that children use their own experiences as a basis for understanding others, often described as "like-me" reasoning (Meltzoff, 2007). Rather than reasoning abstractly about camera systems, children may increasingly recruit their own perceptual experience to infer what another person can see through a mediated interface.

A related process may extend to AR. When perceptual information conflicted, children frequently used bodily movement and physical interaction to evaluate what was real, suggesting reliance not only on perceptual appearance but also on embodied expectations regarding how objects should behave in the physical world. Children may draw on sensorimotor experiences of object permanence, spatial consistency, and contingent interaction to evaluate whether mediated perceptual information aligns with expectations about reality. Such processes are broadly aligned with embodied and constructivist accounts proposing that knowledge of the physical world emerges through action and interaction (Gibson, 1979; Piaget, 1955; Smith & Gasser, 2005).

A similar logic may also help explain social evaluations of artificial agents. In Chapter 3, adults interpreted synchronized musical movement as evidence of warmth, competence, and musical engagement despite limited information regarding the robot's internal capacities. One possibility is that observers recruited embodied models of human musicality and social coordination when interpreting robot behavior. Humans tend to use their own action systems and lived social experiences to interpret others' intentions and engagement, particularly in contexts involving coordinated movement and expressive behavior (Gallese & Goldman, 1998; Meltzoff, 2007; Molnar-Szakacs & Overy, 2006). From this perspective, synchronized movement may

have evoked expectations derived from observers' own experiences of musical engagement, allowing them to interpret coordinated behavior as evidence of perception, intentionality, or social responsiveness. Although the present findings cannot distinguish among embodied simulation, broader social expectations, or attributional heuristics, they suggest that observers may draw on grounded models of human action when interpreting ambiguous agentive cues.

Taken together, the present findings suggest that embodied and self-based knowledge may provide one mechanism supporting RRA by supplying structured expectations about how social and physical systems typically behave. Such mechanisms may operate differently across domains, ranging from direct embodied mapping in video-mediated interaction to broader sensorimotor and social analogies in physical and agentive reasoning. At the same time, observers likely draw on additional sources of knowledge, including causal regularities, conceptual understanding, and social conventions. The self should, therefore, be viewed as a privileged source of embodied experience that may help scaffold interpretation when mediated information is uncertain or ambiguous.

2. RRA as a framework for mediated cognition

The present findings suggest that technology-mediated environments introduce a cognitive challenge that is not fully captured by existing developmental and cognitive theories alone. Prior theoretical frameworks have provided important accounts of how humans reason about perception, representations, and hidden states. Embodied and ecological perspectives emphasize that perception emerges through active engagement with the environment and is grounded in perception-action loops (Gibson, 1979; Smith & Gasser, 2005; Varela et al., 2018). Research on symbolic development and dual representation demonstrates that children gradually learn that representations can stand for reality while differing from it (DeLoache, 2000). Work

on appearance-reality distinctions further shows that children increasingly understand that immediate perceptual appearances may not correspond to underlying states (Flavell et al., 1983). Inferential frameworks, including predictive processing and active inference, characterize cognition as reasoning about hidden causes under uncertainty (Clark, 2019; Friston, 2010). These perspectives offer important foundations for understanding how humans interpret and act upon the world.

My dissertation builds on these traditions while proposing an additional challenge introduced by technology-mediated environments. In many everyday contexts, perceptual information provides relatively stable evidence regarding underlying states. Observers can often assume that visual access corresponds to what another person sees, that perceptual appearance reflects object properties, or that behavioral cues reliably signal intentions and capacities. Technology-mediated environments weaken these assumptions by introducing representational systems that filter, transform, or selectively reveal information. As a result, observers need to determine what hidden state exists, how mediated perceptual signals map onto that state, and whether those signals provide epistemically reliable information.

RRA is proposed as a framework for characterizing this challenge. RRA refers to the process of evaluating how perceptual information maps onto underlying reality and using this relationship to guide action, inference, and judgment. It involves at least three interrelated processes. First, observers must interpret perceptual input as representation, recognizing that what is perceived may not directly correspond to reality but instead reflects information transformed through a representational system. Second, observers must infer hidden states from perceptual evidence, evaluating how available cues map onto underlying social, physical, or agentive structures that are not directly observable. Third, observers must coordinate action to

calibrate representation-reality mappings, using interventions, feedback, and changes in perceptual consequences to evaluate and refine interpretations over time. Successful navigation of mediated environments, therefore, depends not simply on perception or reasoning in isolation, but on dynamically coordinating these processes under conditions of uncertainty.

Unlike traditional accounts of representational understanding, which often emphasize recognizing that representations can differ from reality or stand for something beyond themselves (DeLoache, 1987, 2000; Flavell, 1986b; Perner, 1991), RRA emphasizes the dynamic coordination of perception, inference, and action required to determine what mediated information means during interaction. In other words, RRA is an adaptive process by which observers calibrate the correspondence between mediated representations and underlying social, physical, or agentive states. Critically, RRA does not replace existing theories but instead extends these perspectives by focusing on a problem that becomes especially salient in mediated environments: Determining whether perceptual information itself constitutes reliable evidence and how that information should be used to guide behavior. As such, RRA shifts attention from reasoning about hidden states alone to reasoning about the relationship between perceptual representations and hidden states.

This perspective helps explain why mediated environments often remain challenging even when children possess relevant conceptual knowledge and task performances depend less on their explicit technological knowledge. Traditional developmental accounts frequently ask whether children understand particular concepts, such as perspective, representation, or appearance-reality distinctions. The present findings suggest that mediated environments may additionally require children to coordinate these forms of knowledge under dynamic conditions in which perceptual information must be interpreted in real time. A child may understand that

another person has a different perspective or that appearances can be misleading, yet still struggle to determine how a camera view corresponds to another person's visual access or whether digitally altered features reflect physical reality. From an RRA perspective, developmental adaptation may depend on the ability to dynamically align representations with reality during interaction.

More broadly, RRA aims to provide a useful framework for understanding cognition in increasingly mediated environments. AR, virtual reality, artificial agents, and AI generative systems increasingly shape how people perceive, learn, and interact. These systems often require observers to interpret information that is partial, transformed, or algorithmically generated. As technology-mediated interactions become increasingly common in everyday life, successful adaptation may depend on the ability to evaluate the reliability and meaning of perceptual information across contexts. The developmental findings reported here suggest that this ability may emerge gradually and may depend on opportunities for active calibration through interaction.

3. Design principles for mediated cognition

The present findings suggest that technologies should be designed not only for usability, realism, or engagement (Davis, 1989; Hassenzahl & Tractinsky, 2006; Jig & Lin, 2018; C. Lin et al., 2021; C. Lin, Faas, & Brady, 2017; C. Lin, Faas, Dombrowski, et al., 2017; Norman, 2013; Slater, 2009), but also for how users interpret perceptual information as evidence regarding underlying states. Across video chat, AR, and artificial agents, successful interaction depended on observers' ability to determine how mediated representations corresponded to social, physical, or agentive reality. From the perspective of RRA, technology-mediated interaction often requires users to coordinate perception, action, and inference under conditions in which

perceptual information is indirect, transformed, or incomplete. These inferential demands may be especially consequential for young users, as children gradually learn how mediated signals relate to reality rather than intuitively understanding such mappings.

One implication is that technologies may benefit from supporting representational legibility, or users' ability to understand how perceptual information maps onto underlying states. Rather than assuming that users will naturally infer what mediated information signifies, systems may benefit from making representational mappings more transparent and uncertainty more interpretable. In mediated environments, successful interaction would depend on understanding what that information represents, what it omits, and when it should be treated cautiously. Designs that support users' understanding of how mediated signals are generated, transformed, or constrained may therefore facilitate more effective interpretation and decision making.

A second implication concerns the role of calibration rather than seamlessness in technology design. Many mediated systems prioritize realism, immersion, or frictionless interaction by minimizing ambiguity and perceptual discontinuity. However, the present findings suggest that ambiguity and representational conflict are not necessarily design failures. Under some conditions, uncertainty may function as an epistemic cue that encourages users to explore, test, and evaluate competing interpretations of reality. Particularly in developmental contexts, opportunities for action-based calibration may support learning by helping users determine when mediated information is reliable and when it should be questioned. Human-centered technologies may therefore benefit from supporting active interpretation and hypothesis testing rather than assuming that more seamless perceptual experiences necessarily improve understanding.

Finally, the present findings suggest that mediated systems should support appropriate inference and trust calibration. Humans tend to infer hidden states from behavioral and perceptual cues, including what others can see, whether objects are physically real, and what artificial agents may understand or intend. As technologies increasingly incorporate conversational agents, generative systems, and socially expressive interfaces, users may form assumptions about competence, intentionality, or understanding based on subtle representational cues. From the perspective of RRA, effective design may depend less on maximizing realism or anthropomorphic engagement and more on helping users form appropriately calibrated expectations regarding what systems can and cannot do. As such, trustworthy human-centered technologies may depend on reliable performance as well as supporting accurate alignment between perceptual representations and underlying reality.

4. Limitations and future directions

Several limitations qualify the scope of the present proposal while also suggesting promising directions for future research. Although the dissertation identified a common inferential challenge across video chat, AR, and human-robot interaction, the present studies focused primarily on visually mediated contexts and early childhood, particularly between 3 and 6 years of age. Consequently, the present findings cannot establish whether RRA reflects a shared mechanism across social, physical, and agentive domains or how these processes generalize across development and diverse technological systems. Future work should therefore examine how RRA changes across the lifespan, including older children, adolescents, and adults, as mediated environments become increasingly complex and familiar. One prediction is that children beyond early childhood may become increasingly flexible in revising interpretations when mediated information conflicts with prior expectations, while adults may rely more heavily

on accumulated technological experience or domain-specific assumptions when calibrating perceptual reliability. Future work may also examine populations with atypical developmental trajectories, including autistic children, who may differ in how they interpret social, perceptual, or representational cues in technology-mediated environments. Such comparisons may help clarify whether RRA relies on domain-general inferential processes, embodied self-other mapping, or social-cognitive mechanisms that vary across individuals and developmental pathways.

A second open question concerns the mechanisms through which observers achieve RRA. Although the present findings are broadly consistent with embodied, inferential, and action-based accounts of cognition, the studies were not designed to directly test the processes supporting alignment. The proposal that observers recruit embodied and self-based knowledge to interpret mediated representations, therefore, remains provisional. Future work should distinguish among competing explanations, including predictive inference, embodied self-other mapping, explicit representational reasoning, and strategic problem solving. The present findings also generate several testable predictions. If self-based processes scaffold mediated social reasoning, access to self-relevant information such as self-view should facilitate alignment, whereas disruptions to bodily contingency or self-representation should impair it. If prediction error supports adaptation, increasing representational conflict should encourage exploratory behavior and belief updating when observers treat conflict as epistemically informative. Similarly, if action functions as a mechanism for representational calibration, restricting opportunities for exploration should impair adaptation to uncertain or conflicting mediated information. Computational approaches may further help formalize how observers generate, evaluate, and revise competing hypotheses regarding hidden states under mediation.

Finally, future work should examine not only how RRA succeeds, but also when and why it fails. Increasingly, generative AI, synthetic media, and artificial agents present information that is partial, transformed, or algorithmically generated rather than directly perceived. Under such conditions, failures of alignment may contribute to confusion, misinformation susceptibility, inappropriate trust, or over-attribution of social and agentive qualities to artificial systems. RRA predicts that observers may systematically overweight salient but unreliable mediated cues or underutilize reliable information when representational mappings are unclear. This perspective may also help reinterpret phenomena such as the uncanny valley, in which highly humanlike agents evoke discomfort despite strong perceptual resemblance to humans. From an RRA perspective, uncanny experiences may emerge when perceptual cues strongly imply humanlike underlying states, yet subtle inconsistencies in appearance or behavior disrupt alignment between representation and inferred reality. Understanding how children and adults calibrate trust in mediated information, distinguish meaningful signals from misleading appearances, and revise interpretations under uncertainty may therefore become increasingly important for education, digital literacy, and human-AI interaction. As technologically mediated environments continue to reshape perception and social life, future work on RRA may help clarify not only how humans successfully interpret mediated representations, but also why these interpretations sometimes go wrong.

CONCLUSION

Across video communication, augmented reality, and human-robot interaction, the present dissertation examined how observers determine what perceptual information signifies when relevant social, physical, or agentive states are not directly observable. Successful interaction in these environments increasingly depends on treating perceptual input as evidence, evaluating how representations relate to underlying states, and coordinating perception, representation, action, and inference under uncertainty.

Across developmental and adult contexts, the present findings suggest that representation-reality alignment (RRA) provides a useful framework for characterizing this challenge. Children increasingly used mediated information to infer another person's visual access and to reason about reality when perceptual cues conflicted, while adults used behavioral cues to evaluate possible internal capacities in an artificial agent. Although the present dissertation does not establish a single mechanism underlying these forms of reasoning, the findings are broadly consistent with the idea that mediated environments place shared inferential demands on cognition by requiring observers to evaluate how perceptual information relates to social, physical, and agentive states that cannot be directly observed.

More broadly, the present dissertation suggests that cognition in technology-mediated environments involves evaluating how perceptual representations correspond to reality and what they reveal about underlying states. As digital, augmented, and artificial systems increasingly shape everyday experience, humans may increasingly encounter situations in which perception provides incomplete, transformed, or ambiguous information about the world. Understanding how observers navigate these challenges may therefore become increasingly important for explaining cognition in contemporary mediated environments.

REFERENCES

Clark, A. (2019). *Surfing uncertainty*. Oxford University Press.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *The Mississippi Quarterly*, *13*(3), 319–340.

DeLoache, J. S. (1987). Rapid change in the symbolic functioning of very young children. *Science*, *238*(4833), 1556–1557.

DeLoache, J. S. (2000). Dual representation and young children's use of scale models. *Child Development*, *71*(2), 329–338.

Flavell, J. H. (1986). The development of children's knowledge about the appearance-reality distinction. *The American Psychologist*, *41*(4), 418–425.

Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Development of the appearance--reality distinction. *Cognitive Psychology*, *15*(1), 95–120.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501.

Gibson, J. J. (1979). *Ecological approach to visual perception*. Houghton Mifflin.
Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology*, *25*(2), 91–97.

Jig, K., & Lin, C. (2018). Itchy, Scratchy: A Musical Robot for Low-Income Kid's Piano Practicing. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 347–348.

Lin, C., Faas, T., & Brady, E. (2017). Exploring affection-oriented virtual pet game design strategies in VR attachment, motivations and expectations of users of pet games. *2017 Seventh International Conference*. https://ieeexplore.ieee.org/abstract/document/8273625/

Lin, C., Faas, T., Dombrowski, L., & Brady, E. (2017). Beyond cute: exploring user types and design opportunities of virtual reality pet games. *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, Article Article 1.

Lin, C., Sabanovic, S., Dombrowski, L., Miller, A. D., Brady, E., & MacDorman, K. F. (2021). Parental Acceptance of Children's Storytelling Robots: A Projection of the Uncanny Valley of AI. *Frontiers in Robotics and AI*, *8*, 49.

Meltzoff, A. N. (2007). “Like me”: a foundation for social cognition. *Developmental Science*, *10*(1), 126–134.

Molnar-Szakacs, I., & Overy, K. (2006). Music and mirror neurons: from motion to “e”motion. *Social Cognitive and Affective Neuroscience*, *1*(3), 235–241.

Norman, D. A. (2013). *The design of everyday things*. MIT Press.

Perner, J. (1991). Understanding the representational mind. *Learning, Development, and Conceptual Change.*, *348*. https://psycnet.apa.org/fulltext/1991-97901-000.pdf

Piaget, J. (1955). *The Child’s Construction of Reality*. Routledge & Kegan Paul.

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1535), 3549–3557.

Smith, L., & Gasser, M. (2005). The development of embodied cognition: six lessons from babies. *Artificial Life*, *11*(1-2), 13–29.

Varela, F. J., Thompson, E., & Rosch, E. (2018). *The embodied mind*. MIT Press.