UNIVERSITY OF CALIFORNIA SAN DIEGO

Strategic behavior in social environments

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

in

Experimental Psychology

by

Lauren A. Oey

Committee in charge:

Professor Judith Fan, Co-Chair
Professor Edward Vul, Co-Chair
Professor Leon Bergen
Professor Adena Schachner

2023

The Dissertation of Lauren A. Oey is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It's a funny thing reminiscing about the last five years. Looking at each figure, each argument in this dissertation sparks episodic memories of meetings with advisors filled with laughter, late night work sessions in the graduate housing study rooms, stargazing in the desert, the SPP 2019 sunset dinner at the aquarium overlooking the La Jolla beach, the last pre-pandemic class of developmental proseminar when we used Zoom for the first time, ... Perhaps it was because I started grad school in my early early 20s, or because the pandemic cast some weird time warp, or because I met many of my dearest humans throughout this process. I cannot help but feel that my growth as a person has been tightly coupled with my growth as a scientist. Thank you for all these experiences, and thank you for allowing me this space to be unusually sentimental.

I would like to thank my advisors, Ed Vul and Judy Fan, for training me to be a better computational cognitive scientist, speaker, writer, and critical thinker. Ed, I came to UCSD for your mastery of statistics, savvy argumentation, and astonishing wit. I thought that by the end of five years I will have reached the capacity of what I could learn from another human and will be ready to move on, but that was before I realized that you are in fact superhuman. I will forever treasure your invaluable balance of freedom and guidance in developing my sense of style as a thinker, so thank you. Judy, from your early advice to me as an undergrad summer intern to our boba chats before you started as a professor to now, I am so grateful for your constant presence on my journey. Even before allowing me to join your lab in the eleventh hour of both of our times here at UCSD, you have always been generous in offering me your time and support. You have challenged me to think more critically about my scientific impact and you have taught me to be a better collaborator, so thank you.

This dissertation would not appear as it does without the mentorship of my committee members, Adena Schachner and Leon Bergen. Although projects we worked on together do not appear in this dissertation, I appreciate the knowledge and training you both provided me. Adena—my unofficial third co-chair—our conversations about theory of mind will forever color

my theoretical views. Leon, I promise I will spruce up on linear algebra.

Throughout my time at UCSD, I had a number of informal mentors. I would like to thank Dave Barner, Caren Walker, Lindsey Powell, and Tim Brady for providing feedback on an early version of my dissertation talk. And Vic Ferreira was the first UCSD person I met back at AMLaP 2017, and in interacting with Vic, I knew that UCSD would provide me the supportive environment I was seeking. Other salient formal (and informal) mentors who continually supported me from undergrad to now include Florian, Chigusa, Frank, Linda, Amanda, Wednesday, MH, Robert, and Martin. Lastly, I would like to thank Natalia Vélez who in the last year has helped me to transition my research program in a new direction, and I am excited for what new scientific discoveries we will make together.

A special thank you to Jacob Foster, Erica Cartmill, the rest of the Diverse Intelligences Summer Institute faculty, and my friends that I met there. In the midst of the pandemic, and in the latter half of my graduate career, you reinvigorated and broadened my curiosity in science.

Thank you to my various lab mates and research assistants in the EVul, MaD, and cogtools labs for the various feedback and fun workplace environment you've provided over the years. Way back when interviewing for grad school, I told Ed that I was seeking a "lab mate bromance" to accompany me throughout grad school. I couldn't have asked for a better brother figure and Ph.D. twin than Erik Brockbank.

This dissertation would be incomplete without a tribute to my grad buddy, collaborator, best friend, and pseudo-wife, Holly Huey. Thank you for reading and editing countless versions of my various drafts. And of course, inspiring my research through our various travels, hikes, and coffee shop work sessions. You taught me how to enjoy time off from work. Cheers to many more shared adventures (on the JMT?).

I am grateful for the friends that I met in grad school. My grade-A friends—Alex, Aubrey, Aarthi, Anna—have kept me company while working and relaxing and occupied VIP seats to witness my journey over the years. Also: Minju, Sunyoung, Rose, Elisabeth, Liz, Angus, Lim, Jonas, Mingi, Danbi, Stef, Will, Katja, Isabella, Haleh, Naseem, Janna, Leo, Emily, Srihita, Zoe,

Hannah, Rian, Sunaina, Vicky, Kenny, Hayden, Sara, Alon, Ash, Jeff, Manasvi, Kodi, Althea, Taylor, Kristin, Morgan, Crystal, Jenn, Claire, Judy, Elise, belay partners, skiing buddies, and softball teammates.

Thanks to my family for supporting my decision to pursue a Ph.D.

Thank you to my emotional support dog and emotional support human. My stinky breath Ada, thank you for climbing on me when I am working too late. My wonderful Georgia, thank you for hopping in my passenger seat for the final stretch of my grad school journey, and here's to many future journeys together.

Finally, I must thank my generous funding sources from the NSF Graduate Research Fellowship program and the ACM SIGHPC Computational and Data Science Fellowship.

Chapter 1 is a reprint of material as it appears in DeStefano, I., Oey, L. A., Brockbank, E., & Vul, E. (2021). Integration by parts: Collaboration and topic structure in the CogSci community. *Topics in Cognitive Science*, *13*(2), 399-413. The dissertation author was a primary investigator and co-first author of this paper.

Chapter 2 is a reprint of material as it appears in Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, *152*(2), 346-362. The dissertation author was the primary investigator and author of this paper.

Chapter 3 is a reprint of material that has recently been submitted to *Computational Brain & Behavior* and appears as Oey, L. A., & Vul, E. (under review). Accurate approximations about the truth from literally false messages. An early version of the project was published as a conference proceeding: Oey, L. A., & Vul, E. (2022). Inferring truth from lies. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (pp. 1469-1475). The dissertation author was the primary investigator and author of both these paper.

VITA

2018 Bachelor of Science, Brain and Cognitive Sciences, minor in Computer Science, *magna cum laude*, University of Rochester

2018 Bachelor of Arts, Statistics and Linguistics, University of Rochester

2020 Master of Arts, Experimental Psychology, University of California San Diego

2023 Doctor of Philosophy, Experimental Psychology, University of California San Diego

PUBLICATIONS

Huey, H., Oey, L. A., Lloyd, H. S., & Fan, J. E. (2023). How do communicative goals guide which data visualizations people think are effective? In M. Goldwater, F. Anggoro, B. Hayes, & D. Ong (Eds.), *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, *152*(2), 346–362.

Oey, L. A., & Vul, E. (2022). Inferring truth from lies. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (pp. 1469–1475).

*DeStefano, I., *Oey, L. A., Brockbank, E., & Vul, E. (2021). Integration by parts: Collaboration and topic structure in the CogSci community. *Topics in Cognitive Science*, *13*(2), 399–413.

Oey, L. A., & Vul, E. (2021). Lies are crafted to the audience. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (pp. 791–797).

*Oey, L., *DeStefano, I., Brockbank, E., & Vul, E. (2020). Formalizing interdisciplinary collaboration in the cogsci community. In S. Denison, M. L. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 474–480).

Oey, L. A., Schachner, A., & Vul, E. (2019). Designing good deception: Recursive theory of mind in lying and lie detection. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 897–903).

Oey, L. A., Mollica, F., & Piantadosi, S. T. (2018). Adults use gradient similarity information in compositional rules. In T. T. Rogers, M. Rau, J. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 842–847).

ABSTRACT OF THE DISSERTATION

Strategic behavior in social environments

by

Lauren A. Oey

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2023

Professor Judith Fan, Co-Chair
Professor Edward Vul, Co-Chair

People naturally embed themselves within communities. In doing so, people choose not only whether to cooperate with people in their social environment but *how* to cooperate. Here, I propose that people can produce strategic behaviors as a product of pursuing their goals while predicting how others may interact. Furthermore, individuals' socially motivated behaviors scale up to drive emergent collective behaviors. By examining these behaviors, this dissertation bridges interactions between different social hierarchies—individuals, dyads, and collectives—via recursive social cognitive mechanisms. More specifically, each chapter will address the following questions. Chapter 1 asks how parallel individuals' goals can be used to scaffold how

we understand communities' behavior. Here I analyze how individual scientists' decisions, by combining network and topic space measures, coalesce to determine how integrated the cognitive science community has become in the last two decades. Chapter 2 tests how individuals deceive with lies via recursive social reasoning and strategic planning to dodge others' enforcement. Chapter 3 considers how enforcers adapt their expectations about what interpretation to take away from messages when they suspect defectors. This shift in how people interpret messages can then lead to runaway effects at the collective level, such as imprecise approximations about the truth and cooperative speakers also being induced to lie for their messages to be understood accurately. Overall, this work lays the foundation toward a unifying theory of people's and collective's behaviors arising from individualistic goals in social environments.

Chapter 0

Introduction

People naturally embed themselves within communities to belong within a group identity, to retain and innovate knowledge, to institute laws and conventions, and to design and dismantle establishments. By joining social environments, people agree to partake in and influence these shared collective behaviors. In turn, people are rewarded with cultural affordances. They inherit evolved solutions to problems or ambiguities experienced by peers, elders, and previous generations in the form of cultural knowledge and norms. Moreover, they are bestowed the trust and collaboration of fellow community members. Equipped with this social contract (Rousseau, 1916), human communities may develop technologies and institutions that no single individual could create on their own (Henrich, 2015; Tomasello, 2000). Following this logic, communities would perhaps benefit if they marshaled people to prioritize the goals of the community and surrender individual goals that conflict. Instead, people retain individual freedoms to choose how they balance their self-interests with societal goals (Arrow, 1974).

Economists and political philosophers have long debated whether individuals require top-down governed rules to cooperate within collectives. Sometimes, as in an Invisible Hand mechanism (Smith, 1776), individuals (or consumers) may follow their own goals, which so happens to culminate in the best outcome for collectives (or economies). Other times, individual goals seem to hinder the collective. For example, people may be tempted to discard waste along hiking trails to relieve themselves of carrying around trash. However, if everyone littered and

fail to account that others are doing the same, then trails will be polluted with hazards that spoil nature for fellow outdoor enthusiasts and harm wildlife. Similarly, a fisherman seeking to maximize their earnings may over-fish relative to their fair share, and more broadly people may over-consume limited *public goods*, resulting in a scarcity for the collective (Hardin, 1968). An individual taking one unit "more" from the pool of goods causes minimal harm. But when many or everyone shares that same assumption, the consequences for the collective can become dire (Kant, 1785; S. Levine et al., 2020). In each individual's best interest, they may want others in their social environment to cooperatively share, but for themselves, doing so forgoes potential personal gains or even bears costs. Others in their social environment may feel the same. Here lies the social dilemma: ***How do people strategically behave when pursuing their own goals while embedded within their local social environments?***

In this introduction, I will first address why collectives, individuals, and simple dyadic interactions—on their own—fall short of explaining how people choose to cooperate or defect. I will then consider strategic behaviors that people do in everyday life that skirt the boundaries of cooperation and defection. I argue that these strategic behaviors require people to use recursive social reasoning across social hierarchies. Finally, I explain how the research in my dissertation contributes to solving this social dilemma. Specifically, through computational and empirical methods, I show how individuals produce strategic behaviors in pursuit of their own goals within their local social environment, and how individual goals within social environments can ultimately drive emergent collective behaviors and norms.

0.1 Collectives, individuals, and dyadic interactions cannot on their own resolve the conflict between cooperation and defection

What role do collectives play to support cooperation from individuals? Collectives can top-down regulate individuals' behavior to discourage defection (Hilbe et al., 2014; Hobbes, 1651; O'Gorman et al., 2009; Yamagishi, 1986). By creating centralized authorities and laws,

collectives can allocate labor to enforce cooperation. Alternatively, real-world communities are characterized by social network structures that give rise to bottom-up emergent behaviors (e.g., Apicella et al., 2012). Computer simulations reveal that select graph features within communities facilitate the spread and maintenance of cooperation throughout the network (Centola, 2013; Cohen et al., 2001; Lieberman et al., 2005; McAvoy et al., 2020; Taylor et al., 2007). Thus human populations, trading off top-down and bottom-up mechanisms, help reduce tensions between cooperative collectives and defective individuals (Sigmund et al., 2010). Importantly, collective mechanisms alone cannot explain human cooperation. First, many everyday decisions, like lying on dating apps, represent low impact defections that are not problematic enough to be worthy of regulation by authorities. Second, individuals have the option to resist authorities in these everyday decisions, but they instead *choose* to be cooperative subjects. Third, an individual's location within a network may augment the ease with which one can defect (e.g., a person who exists in social environment with stricter norms around lying may be more resistant to lie), but the final decision of when or how to lie remains a freedom of the individual. Therefore, human cooperation must be understood not only through institutions, but also through individuals.

What role do individuals play as cooperative subjects? While populations have thus far been characterized by the structure of their social networks, let us consider the nodes that compose networks: individuals. Economists have traditionally theorized that individuals will choose to cooperate when the benefit to their self-interests outweigh the costs. On the contrary, across real-world situations and numerous behavioral economic experiments, people do not simply navigate the world by being selfish (Camerer, 2014). In fact, as a species, humans tend to display stable *prosocial* personality traits (Zaki & Mitchell, 2013), even incurring costs to benefit others' self-interests. Humans automatically prefer to be altruistic (De Waal, 2008), fair (Thaler, 1988), honest (Abeler et al., 2019; Capraro, Schulz, et al., 2019; Shalvi et al., 2012; Suchotzki et al., 2017), and trusting (Bond & DePaulo, 2006; Brashier & Marsh, 2020; T. R. Levine, 2014). The assumption that people default to prosocial behaviors pervades the current literature, which now focuses on the social and environmental factors that drive people to deviate

from cooperation. What demographics, cultural, or other idiosynchratic traits predispose people to defect or cooperate (Engelmann et al., 2019; Henrich et al., 2001; House et al., 2013; Rand et al., 2016)? How can policies prevent people from deviating from baseline traits (Capraro, Jagfeld, et al., 2019; Kraft-Todd et al., 2015)? And, ultimately, how do prosocial traits feed into the "success" of individuals, populations, species (Hare, 2017; Henrich, 2015)? While people may have a bias to be prosocial, the focus on stable traits neglects that people can and do adapt to antisocial environments. For example, trust endows people to learn about the latest scientific advancements, but gullibility leads people astray in phishing scams. Thus, trait explanations assume monotonic behavior, but miss that people change how prosocial they choose to be, based on the situation around them. Thus, people ought to be situationally-aware decision makers that make active decisions about when to shift from prosocial tendencies and instead be skeptical.

Scaled between nodes and network structures, let us consider the links between nodes. Networks can be decomposed into dyadic actions between individuals. Which links exist, and which do not, define the social network structure. *How do the actions between individuals support cooperation?*

Evolutionary game theory has shown that agents can effectively discourage defection by implementing select strategies for acting on others in repeated interactions. People mimic each other. By reciprocating the same action back (tit-for-tat) or reflexively changing behaviors based on the outcome (win-stay, lose-shift), people select a commonly used, and evolutionarily stable, strategy (Axelrod & Hamilton, 1981; Fehr & Fischbacher, 2003, 2004a; M. Nowak & Sigmund, 1993). If one person is the recipient of another's altruism, that person feels compelled to return the altruism (Trivers, 1971). Vice versa, people who feel burned by another's defection may retaliate by withholding cooperation or directly punishing the defector (Clutton-Brock & Parker, 1995). Mimicking in this manner also serves as a form of simple reinforcement learning. Defectors are punished for negative behaviors and cooperators are rewarded for positive behaviors.

The impact of defection extend beyond just the direct victims of the action. Defectors

are not only negatively perceived because they violate norms or laws, but they threaten to undermine cooperation across the community (Fehr & Fischbacher, 2003; Sarkadi et al., 2021), thus motivating the rise of (decentralized) regulation. If some people aren't cooperative, third parties adopt community goals to monitor, reinforce desirable, or punish undesirable behavior (Boyd & Mathew, 2021; Fehr & Fischbacher, 2004b; M. A. Nowak & Sigmund, 2005). Prosocial actions also boosts cooperation in third parties. Witnessing one good Samaritan, people may choose to perform a good deed for another person. In this way, cooperative behaviors can contagiously spread from neighbor to neighbor in a community, by simply mimicking observed cooperative behaviors (Fowler & Christakis, 2010).

Mimicking and simple reinforcement learning are simple actions that have useful but limited explanatory power to explain the widespread adoption of cooperation. What they lack are the ability to explain how people robustly adapt to different social communities and situations. Instead we might need social cognitive mechanisms, which allow us to judge whether others intend to cooperate or defect or whether others expect us to do so, and adapt to different social scenarios. Consider the following cases that require more robust mechanisms than what mimicking and simple reinforcement have to offer. First, the effectiveness of reinforcements to block defection demands that people be more omniscient defector detectors than they really are. In real world situations, people systematically fail to detect defections. For example, when attempting to detect verbal lies, people are systematically biased to believe statements as true (Bond & DePaulo, 2006), in part because there exists limited diagnostic information that can reliably discriminate lies from truths (Street, 2015). Or even when people do detect defections, they enforce under select conditions, e.g., it will personally benefit their reputation (Pedersen et al., 2018). Second, both copying and reinforcement cannot capture the rich ways that people decide when and how to prioritize their individual goals over collective goals by defecting. People are more willing to violate cooperative norms in certain contexts and not in others, e.g. on dating apps people lie more frequently about height or age than marital status (Toma et al., 2008). Or people can defect in ways that strategically bypass punishment. People prefer to mislead or

withhold information over outright lying (Montague et al., 2011; Ransom et al., 2019; Rogers et al., 2017), perhaps because they are conventionally perceived as less reprehensible (Schauer & Zeckhauser, 2007). Even children plan clever loophole methods to be non-compliant, yet avoid punishment (Bridgers et al., 2021). Third, when people are detected and punished, corrections do not always work, in fact they can backfire, causing people to further deviate from collective goals (Mosleh et al., 2021; Nyhan & Reifler, 2010). All these counterpoints require that people select how to act by predicting and avoiding others' negative reception through social cognitive mechanisms.

0.2 Social cognitive mechanisms bind interactions across collectives, individuals, and dyads

Human social cognitive mechanisms underlie how people select actions that trade off cooperating or defecting across varying contexts. When observing other people's actions, people can rapidly intuit others' beliefs and goals (Baker et al., 2017; Jara-Ettinger et al., 2016). The objective of these social inferences is to reverse engineer abstract, structured mental models of how others operate (Carlson et al., 2022; Kleiman-Weiner et al., 2016; Kleiman-Weiner et al., 2017; Ullman et al., 2009; Wu et al., 2021). Harnessing these inferred models to simulate how others will behave, people can adaptively plan in ways that reactively *and proactively* respond to others' goals, beliefs, and behaviors (Brockbank & Vul, 2021; Ho et al., 2022).

Social situations are often *interactive*. People align their representations (Pickering & Garrod, 2004), negotiate common ground (Clark & Wilkes-Gibbs, 1986), and coordinate by communicating words (Clark, 1996; Wilkes-Gibbs & Clark, 1992), gestures (Goldin-Meadow, 1999), demonstrations (Ho et al., 2021), examples (Shafto et al., 2014), drawings (Fan et al., 2020), data visualizations (Huey et al., 2023), or without any signals at all (Schelling, 1960). In communicating, people orient what they say to others so that others will easily understand what was said (Fussell & Krauss, 1989, 1992). People favor saying what is relevant or informative

over what is useless or obvious to what others already know (Bannard et al., 2017; Bergey et al., 2020; Degen et al., 2020; Sperber & Wilson, 1986), and listeners extract meaning by assuming messages are intended to be informative (Frank & Goodman, 2012, 2014; Grice, 1975). Furthermore, people are inclined to send signals that exhibit their communicative intent (Ho et al., 2019; Sarin et al., 2021; Scott-Phillips et al., 2009) and audiences read signals as communicative intent (Ho et al., 2017). Importantly, people perceive, plan, act, and provide feedback to each other.

When both or multiple people simulate each other as social reasoning agents, the result is a recursively iterative process: "I think about you, thinking about me, thinking about you, etc.", in which people attempt to socially reason "one level ahead" of others (Camerer et al., 2004). Recursive social reasoning has productively been used to explain how people engage in various cooperative communicative interactions, like how people teach and learn effectively from others (Bonawitz et al., 2011; Gweon, 2021; Shafto et al., 2014) and speak with and comprehend conversational partners in language (Franke, 2013; Goodman & Frank, 2016; Hawkins et al., 2019). In these purely cooperative interactions, people converge on choosing actions that optimize for both the individual and others within very few levels of recursion.

Many real world interactions, however, are not purely cooperative. In everyday life, people navigate thornier interactions with risks for miscommunication, such as dodging faux pas, meddling in gossip, telling white lies, and debating with someone to change their views. Here, individuals' goals conflict with others (the collective)—consider how a verbal spar can slip into an insult brawl, completely undermining cooperation. In these socially risky interactions, it is critical that people can accurately represent another's mind to avoid exacerbating the interaction. Importantly, carving out a true representation of another's mind relies on using what they say and do to learn about what they know and think.

When building a model of someone else's mind, people share a latent assumption: people can reliably trust that what another person says is a *true* analog to what that other person thinks. This assumption follows from a broader assumption that other people try to communicate

cooperatively, such as by being informative and honest (Grice, 1975). The assumption is further reinforced by members in the community speaking truthfully. In fact, David Lewis (Lewis, 1969, 1975) argues that the dyadic and community conventions around speaking truthfully and listening trustfully permit language to exist and to be understood at all. In this way, truth in communication is a *public good*. Everyone reaps the benefit that communication is generally perceived as and often is cooperative. Knowing that the community contributes to the collective goal to maintain truthfulness in communication, people can expect others to trust that what they say as what they verbatim think and intend to be true. Thus, communication can typically be efficient without needing to provide additional signals to verify the veridicality of a statement.

Like all public goods, truthful communication can be taken advantage of by individual liars. Even absent intentional liars, speakers may mistakenly provide erroneous or incomplete information, and messages may become distorted in noisy communication channels. When false or noisy information is present, people should exercise caution to protect oneself from being led astray (Sperber et al., 2010). In fact, people can diagnose intentions to help or hinder (Kleiman-Weiner et al., 2016; Ullman et al., 2009), even from infancy (Hamlin et al., 2007). By five to nine, children develop a sensitivity to when others are less than fully informative or honest. Children recognize what information they think others have access to, what information they think is omitted, and whether honest or dishonest intentions drive communication (Gweon et al., 2014; Kominsky et al., 2016; Mascaro & Sperber, 2009; Shafto et al., 2012). In adulthood, people can adjust how they integrate others' provided erroneous or imperfect information into their own decision making process (Ronfard & Lane, 2019; Vélez & Gweon, 2019). While listeners may use smart strategies to detect and interpret dishonesty, smart speakers may also be wary of vigilant listeners. Just as cooperative speakers signal their intentions, adversarial speakers occlude their intentions (Strouse et al., 2018), such as by building in plausible deniability about what their intention could have been (Lee & Pinker, 2010; Pinker et al., 2008). Thus, the balancing of individual and collective goals introduces an arms race between socially strategic speakers and listeners. Here, recursive social reasoning may play a key role in equipping people

with the ability to resist defectors that violate truthful communication.

0.3 Current Directions

In this dissertation, I propose that people produce strategic behaviors, as a product of pursuing their goals while being embedded within social environments, and these socially motivated behaviors can drive collective behaviors. Across three chapters, this dissertation bridges interactions between different social hierarchies—individuals, dyads, and collectives—via recursive social cognitive mechanisms. Grounding these three social levels in their interactions have already begun to inspire our understanding of how collective behaviors evolve, such as in convention formation (Hawkins et al., 2019). The novel contribution of my work to this growing literature is that strategic evasive defection interacting with strategic enforcement can spawn new collective behaviors and norms to evolve. More specifically, I will address the following questions in each chapter: (1) how can parallel individuals' goals be used to scaffold how we understand communities' behavior, (2) how do individuals socially reason and strategically plan to dodge others' enforcement, and (3) how do enforcers adapt to expectations about defectors, which can then lead to runaway effects at the collective level?

In Chapter 1, I analyze how parallel individual decisions coalesce to drive a community's integrative behaviors (*DeStefano, *Oey et al., 2021). I explore this collective behavior using the cognitive science community as a case study. Since its inception, cognitive science has collectively aspired to be an interdisciplinary field. However, recent metascientific papers have debated whether we have consummated this aspiration (Contreras Kallens et al., 2022; Gray, 2019; Núñez et al., 2019, 2020). In this chapter, I tackle the question of whether cognitive scientific is an *integrated* community, one key attribute of interdisciplinarity. For my overarching framework to quantify community integration, I borrow a core tenet of cognitive science: agents perform actions that influence the behavior of systems (Goldstone, 2019). Each individual cognitive scientist is a goal-oriented agent that seeks to form co-authorship relationships with

others and study research topics that guide scientific progress. Scientists' parallel decisions amalgamate to form the interconnectedness of the cognitive science network. Critically I apply large-scale data science techniques to measure how the structure of scientific co-authorship networks and the topic space become more integrated or pocketed over time. I find that the network and topic space measures both point toward increasing integration in cognitive science. More broadly, this study highlights the value of understanding collective behaviors through the lens of bridging individuals, their goals to link up, and changes to the resulting social network structure.

The individual goals I consider in Chapter 1 are broadly cooperative: scientists link up to blend varying expertise and to form mentorships. But there also exist competitive goals in science: to forge a distinct research program relative to other scientists, to bridge research areas or trailblaze new unexplored areas before others can, etc. Individuals that strive to discover new inexplicable, paradigm-shifting puzzles might lose sight of the importance of scientific validity and replicability. As a result, these individuals derail from the overall collective goal for scientific progress or even mislead the field into exploring dead ends (e.g., GOFAI). Thus other scientists should be vigilant to mismatched goals. In the second chapter, I examine how vigilance to mismatched goals influences how people detect defection, within the domain of dyadic communication. Furthermore, individuals who interactively reason about others may strategically select actions that avoid detection.

In Chapter 2, I test how individuals may strategically flout cooperation by designing actions to avoid detection when lying in everyday communication (Oey et al., 2023). Across communication channels, communities seek to maintain norms around telling the truth. However, individuals have the free will to defect, but they must be strategic about what lies they say or else risk being caught. Social cognitive mechanisms that guide audiences to detect lies are the same mechanisms that allow speakers to choose lies to go undetected. Applying a first principles approach, I introduce a Bayesian model of interactive and recursive social reasoning to explain lying behaviors. Speakers seek to exaggerate claims to benefit themselves (e.g., "I have published

12 high-impact research papers!"), but they want to be compliant with collective goals to appear honest. Meanwhile, listeners seek to detect lies (e.g., "That must be false—there is no way you published *that* many high-impact papers"), but they want to be compliant with collective goals to trust when statements are indeed true. As a starting point, I focus on dyads as a simpler model for multi-agent enforcement. Crucially, recursive social cognitive mechanisms better explain how people lie and detect lies, compared to simpler non-social heuristics. In the third chapter, we consider potential downstream consequences for speakers and listeners who recursively reason about lying.

Thus, in Chapter 3, I consider how listeners adjust their social expectations about how to interpret messages believes to be lies, and what downstream effects this would have on community-wide communication systems (Oey & Vul, 2022, under review). In the previous chapter, speakers manipulate messages along the extremity of the claim, and listeners detect lies as true or false. In this chapter, I consider listeners as more advanced agents that seek to glean meaning from messages (e.g., "Although you said you published 12 high-impact papers, in actuality you probably published 3."). Optimistically, I show that listeners can and do go beyond detecting lies to interpret what is the actual truth by considering the goals and costs faced by speakers. More pessimistically, the introduction of biased messages at all can undermine how communication systems map messages onto meaning, as shown through computer simulations. Even in ideal situations in which listeners are perfectly tuned to the goals and costs of speakers, they extract accurate *but imprecise* approximations to the truth. Furthermore, when listeners generalize their skepticism across speakers, cooperative speakers must also lie to have their messages accurately understood. This chapter introduces a potential evolutionary account of why certain communication channels, like letters of recommendation, are systematically exaggerated. More broadly, this work scales up communication from a focus on individuals and dyads to wider consequences for populations.

Altogether, this dissertation asks how people strategically pursue their goals within social environments by applying interactive and recursive social reasoning. I examine social

decision making and communication across varying social group levels—collectives to dyads to individuals—and their interactions. I apply data scientific methods, computational models, and behavioral studies to show that recursive social reasoning robustly explains various real-world lying behaviors. Overall, this work lays the foundation for a unifying theory of individualistic goals in social environments.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115–1153.

Apicella, C. L., Marlowe, F. W., Fowler, J. H., & Christakis, N. A. (2012). Social networks and cooperation in hunter-gatherers. *Nature*, *481*(7382), 497–501.

Arrow, K. J. (1974). *The Limits of Organization*. WW Norton & Company.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *SWcience*, *211*(4489), 1390–1396.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires, and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1–10.

Bannard, C., Rosner, M., & Matthews, D. (2017). What's worth talking about? information theory reveals how children balance informativeness and ease of production. *Psychological Science*, *28*(7), 954–966.

Bergey, C., Morris, B., & Yurovsky, D. (2020). Children hear more about what is atypical than what is typical (S. Denison, M. L. Mack, Y. Xu, & B. C. Armstrong, Eds.), 501–507.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234.

Boyd, R., & Mathew, S. (2021). Arbitration supports reciprocity when there are frequent perception errors. *Nature Human Behaviour*, *5*(5), 596–603.

Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, *71*, 499–515.

Bridgers, S., Schulz, L. E., & Ullman, T. D. (2021). Loopholes, a window into value alignment and the learning of meaning. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (pp. 126–132).

Brockbank, E., & Vul, E. (2021). Formalizing opponent modeling with the rock, paper, scissors game. *Games*, *12*(3), 70.

Camerer, C. F. (2014). Behavioral economics. *Current Biology*, *24*(18), R867–R871.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, *119*(3), 861–898.

Capraro, V., Jagfeld, G., Klein, R., Mul, M., & van de Pol, I. (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. *Scientific Reports*, *9*(1), 1–11.

Capraro, V., Schulz, J., & Rand, D. G. (2019). Time pressure and honesty in a deception game. *Journal of Behavioral and Experimental Economics*, *79*, 93–99.

Carlson, R. W., Bigman, Y. E., Gray, K., Ferguson, M. J., & Crockett, M. (2022). How inferred motives shape moral judgements. *Nature Reviews Psychology*, *1*(8), 468–478.

Centola, D. M. (2013). Homophily, networks, and critical mass: Solving the start-up problem in large group collective action. *Rationality and Society*, *25*(1), 3–40.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*(6511), 209–216.

Cohen, M. D., Riolo, R. L., & Axelrod, R. (2001). The role of social structure in the maintenance of cooperative regimes. *Rationality and Society*, *13*(1), 5–32.

Contreras Kallens, P., Dale, R., & Christiansen, M. H. (2022). Quantifying interdisciplinarity in cognitive science and beyond. *Topics in Cognitive Science*, *14*(3), 634–645.

De Waal, F. B. M. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology*, *59*, 279–300.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591.

*DeStefano, I., *Oey, L. A., Brockbank, E., & Vul, E. (2021). Integration by parts: Collaboration and topic structure in the CogSci community. *Topics in Cognitive Science*, *13*(2), 399–413.

Engelmann, J. B., Schmid, B., De Dreu, C. K., Chumbley, J., & Fehr, E. (2019). On the psychology and economics of antisocial personality. *Proceedings of the National Academy of Sciences*, *116*(26), 12781–12786.

Fan, J. E., Hawkins, R. D., Wu, M., & Goodman, N. D. (2020). Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, *3*, 86–101.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791.

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190.

Fehr, E., & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87.

Fowler, J. H., & Christakis, N. A. (2010). Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, *107*(12), 5334–5338.

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

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Franke, M. (2013). Game theoretic pragmatics. *Philosophy Compass*, *8*(3), 269–284.

Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, *25*(3), 203–219.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology*, *62*(3), 378.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, *3*(11), 419–429.

Goldstone, R. L. (2019). Becoming cognitive science. *Topics in Cognitive Science*, *11*(4), 902–913.

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

Gray, W. D. (2019). Welcome to Cognitive Science: The once and future multidisciplinary society. *Topics in Cognitive Science*, *11*(4), 838–844.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics Vol. 3: Speech Acts* (pp. 64–75). Academic Press.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910.

Gweon, H., Pelton, H., Konopka, J. A., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when teachers are under-informative. *Cognition*, *132*(3), 335–341.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559.

Hardin, G. (1968). The tragedy of the commons. *Science*, *162*(3859), 1243–1248.

Hare, B. (2017). Survival of the friendliest: *Homo sapiens* evolved via selection for prosociality. *Annual Review of Psychology*, *68*, 155–186.

Hawkins, R. X., Goodman, N. D., & Goldstone, R. L. (2019). The emergence of social norms and conventions. *Trends in Cognitive Sciences*, *23*(2), 158–169.

Henrich, J. (2015). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton University Press.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of Homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–78.

Hilbe, C., Traulsen, A., Röhl, T., & Milinski, M. (2014). Democratic decisions establish stable authorities that overcome the paradox of second-order punishment. *Proceedings of the National Academy of Sciences*, *111*(2), 752–756.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, *148*(3), 520–549.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2021). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*, *150*(11), 2246.

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition*, *167*, 91–106.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*, *26*(11), 959–971.

Hobbes, T. (1651). *Leviathan*. Simon & Schuster.

House, B. R., Silk, J. B., Henrich, J., Barrett, H. C., Scelza, B. A., Boyette, A. H., Hewlett, B. S., McElreath, R., & Laurence, S. (2013). Ontogeny of prosocial behavior across diverse societies. *Proceedings of the National Academy of Sciences*, *110*(36), 14586–14591.

Huey, H., Oey, L. A., Lloyd, H. S., & Fan, J. E. (2023). How do communicative goals guide which data visualizations people think are effective? In M. Goldwater, F. Anggoro, B. Hayes, & D. Ong (Eds.), *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(10), 589–604.

Kant, I. (1785). *Groundwork of the Metaphysics of Morals*. Cambridge University Press.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction. In A. Papafragou, D. J. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 1679–1684).

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123.

Kominsky, J. F., Langthorne, P., & Keil, F. C. (2016). The better part of not knowing: Virtuous ignorance. *Developmental Psychology*, *52*(1), 31.

Kraft-Todd, G., Yoeli, E., Bhanot, S., & Rand, D. (2015). Promoting cooperation in the field. *Current Opinion in Behavioral Sciences*, *3*, 96–101.

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, *117*(3), 785–807.

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, *117*(42), 26158–26169.

Levine, T. R. (2014). Truth-Default-Theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*(4), 1–15.

Lewis, D. (1969). *Convention: A Philosophical Study*. John Wiley & Sons.

Lewis, D. (1975). Languages and language.

Lieberman, E., Hauert, C., & Nowak, M. A. (2005). Evolutionary dynamics on graphs. *Nature*, *433*(7023), 312–316.

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–380.

McAvoy, A., Allen, B., & Nowak, M. A. (2020). Social goods dilemmas in heterogeneous societies. *Nature Human Behaviour*, *4*(8), 819–831.

Montague, R., Navarro, D. J., Perfors, A., Warner, R., & Shafto, P. (2011). To catch a liar: The effects of truthful and deceptive testimony on inferential learning. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 1312–1317).

Mosleh, M., Martel, C., Eckles, D., & Rand, D. (2021). Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, *364*, 56–58.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298.

Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., & Semenuks, A. (2019). What happened to cognitive science? *Nature Human Behaviour*, *3*(8), 782–791.

Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., & Semenuks, A. (2020). For the sciences they are a-changin': A response to commentaries on Núñez et al.'s (2019) "What happened to cognitive science?" *Topics in Cognitive Science*, *12*(3), 790–803.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, *152*(2), 346–362.

Oey, L. A., & Vul, E. (2022). Inferring truth from lies. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (pp. 1469–1475).

Oey, L. A., & Vul, E. (under review). Accurate approximations about the truth from literally false messages.

O'Gorman, R., Henrich, J., & Van Vugt, M. (2009). Constraining free riding in public goods games: Designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1655), 323–329.

Pedersen, E. J., McAuliffe, W. H., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General*, *147*(4), 514.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190.

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of sciences*, *105*(3), 833–838.

Rand, D. G., Brescoll, V. L., Everett, J. A., Capraro, V., & Barcelo, H. (2016). Social heuristics and social roles: Intuition favors altruism for women but not for men. *Journal of Experimental Psychology: General*, *145*(4), 389.

Ransom, K., Voorspoels, W., Navarro, D. J., & Perfors, A. (2019). Where the truth lies: How sampling implications drive deception without lying. *PsyArXiv*.

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead other. *Journal of Personality and Social Psychology*, *112*(3), 456–473.

Ronfard, S., & Lane, J. D. (2019). Children's and adults' epistemic trust in and impressions of inaccurate informants. *Journal of Experimental Child Psychology*, *188*, 104662.

Rousseau, J.-J. (1916). *The social contract: Or, principles of political right*. G. Allen & Unwin, Limited [1916].

Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, 104544.

Sarkadi, Ş., Rutherford, A., McBurney, P., Parsons, S., & Rahwan, I. (2021). The evolution of deception. *Royal Society Open Science*, *8*(9), 201032.

Schauer, F., & Zeckhauser, R. J. (2007). Paltering.

Schelling, T. C. (1960). *The Strategy of Conflict*. Harvard University.

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226–233.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, *15*(3), 436–447.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justification). *Psychological Science*, *23*(10), 1264–1270.

Sigmund, K., De Silva, H., Traulsen, A., & Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature*, *466*(7308), 861–863.

Smith, A. (1776). *The Wealth of Nations*. W. Strahan; T. Cadell.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition* (Vol. 142). Citeseer.

Street, C. N. (2015). ALIED: Humans as adaptive lie detectors. *Journal of Applied Research in Memory and Cognition*, *4*(4), 335–343.

Strouse, D., Kleiman-Weiner, M., Tenenbaum, J., Botvinick, M., & Schwab, D. J. (2018). Learning to share and hide intentions using information regularization. *Advances in Neural Information Processing Systems*, 10249–10259.

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428–453.

Taylor, P. D., Day, T., & Wild, G. (2007). Evolution of cooperation in a finite homogeneous graph. *Nature*, *447*(7143), 469–472.

Thaler, R. H. (1988). Anomalies: The ultimatum game. *Journal of Economic Perspectives*, *2*(4), 195–206.

Toma, C. L., Hancock, J. T., & Ellison, N. B. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin*, *34*(8), 1023–1036.

Tomasello, M. (2000). *The Cultural Origins of Human Cognition*. Harvard University Press.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems*, *22*, 1874–1882.

Vélez, N., & Gweon, H. (2019). Integrating incomplete information with imperfect advice. *Topics in Cognitive Science*, *11*(2), 299–315.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*(2), 183–194.

Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., & Kleiman-Weiner, M. (2021). Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, *13*(2), 414–432.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*(1), 110.

Zaki, J., & Mitchell, J. P. (2013). Intuitive prosociality. *Current Directions in Psychological Science*, *22*(6), 466–470.

Chapter 1

Integration by parts: Collaboration and topic structure in the CogSci community

Is cognitive science interdisciplinary or multidisciplinary? We contribute to this debate by examining the authorship structure and topic similarity of contributions to the Cognitive Science Society from 2000 to 2019. Our analysis focuses on graph theoretic features of the co-authorship network—edge density, transitivity, and maximum subgraph size—as well as clustering within the space of scientific topics. We also combine structural and semantic information with an analysis of how authors choose their collaborators based on their interests and prior collaborations. We compare findings from CogSci to abstracts from the Vision Science Society over the same time frame and validate our approach by predicting new collaborations in the 2020 CogSci proceedings. Our results suggest that collaboration across authors and topics within cognitive science has become increasingly integrated in the last 19 years. More broadly, we argue that a formal quantitative approach which combines structural co-authorship information and semantic topic analysis provides inroads to questions about the level of interdisciplinary collaboration in a scientific community.

Keywords: co-authorship networks, topic modeling, interdisciplinarity, multidisciplinarity, scientometrics

1.1 Introduction

Since its foundation, the Cognitive Science Society sought to unify various disciplines of study under one interdisciplinary research field. Recently, criticism of the success of this mission has sparked debate about whether cognitive science, in its current form, is fundamentally multidisciplinary rather than interdisciplinary (Gray, 2019; Núñez et al., 2019; Schunn et al., 1998). The distinction between these community structures is subtle, making any claims favoring one or the other difficult to evaluate. Broadly, the debate centers on the idea that a research community is more *multidisciplinary* if collaborations happen mostly within small groups and there is greater topical isolation of each group from the rest. On the other hand, a more *interdisciplinary* research community will show fewer isolated groups and less separation of research interests across groups. Researchers hoping to promote progress in a field might therefore strive for a more interdisciplinary, rather than multidisciplinary, approach.

But how do we measure interdisciplinarity in a way that captures meaningful differences within diverse communities? Currently, there is no consensus on a single measure that best aligns with this abstract concept. Previous studies quantified interdisciplinarity by looking at the publication record in journals associated with a given discipline. Some of these studies have examined the distribution of journals cited (Goldstone & Leydesdorff, 2006; Núñez et al., 2019; Porter et al., 2007), the citation networks (Rafols & Meyer, 2010), and the journals that authors previously published in (Bergmann et al., 2017). But this earlier research aiming to quantify interdisciplinarity was primarily targeted at the categorization of disciplines. These measures are subject to inconsistencies across classification systems, leading to variable conclusions (Wagner et al., 2011). Others have used departmental affiliation and educational background (Núñez et al., 2019; Schunn et al., 1998), but research interests often shift over the course of a lifetime, which makes the affiliation label a transient indicator (Porter et al., 2007).

Recent efforts to measure interdisciplinarity or characterize the level of collaboration in a field have sought to address these challenges by incorporating more data-rich, bottom-up

measures. For example, the *contents* of scientific work in a number of fields outside cognitive science have been described using text-based clustering (Gowanlock & Gazan, 2013), word co-occurrence (Ravikumar et al., 2015), semantic structural analysis (Parinov & Kogalovsky, 2014), and topic modeling (Nichols, 2014). Further, the *structural* properties of research collaboration have been described using network analysis tools applied to publication in diverse scientific fields (Barabási et al., 2002; Newman, 2001, 2004), in management and organizational research (Acedo et al., 2006), and in international collaborations (Wagner & Leyesdorff, 2005). These measures offer the ability to characterize work in a field without relying on the manual assignment of authors or publications to particular disciplines.

Based on this work, what conclusions can be drawn about cognitive science specifically? A recent comprehensive attempt to assess whether cognitive science reached the interdisciplinary status it aspired to, comes from Núñez et al. (2019). The authors combine bibliometric indicators– the affiliation of authors in the journal *Cognitive Science* and the disciplines of journals cited therein—as well as socio-institutional ones: the doctoral training of faculty in cognitive science departments and the coursework requirements of cognitive science undergraduate cores. Both of these latter measures are already constrained by the few institutions that offer undergraduate training or have separate departments in cognitive science at all. Núñez et al. (2019) conclude that there is an imbalanced contribution of the constituent disciplines to cognitive science, suggesting that cognitive science remains premature in its efforts to forge a coherent interdisciplinary field. The results sparked controversy and a range of responses (see overview in Gray, 2019 and Núñez et al., 2020), both theoretical and empirical. Many of these addressed the inherent challenges of measuring interdisciplinarity, noting for example that an author's departmental affiliation provides at best "a useful proxy for a scholar's background" (Bender, 2019). Thus, the discussion about the level of interdisciplinary work in cognitive science may benefit from more fine-grained measures of author affiliations and research areas.

In the current work, we aim to contribute to the debate over interdisciplinarity in cognitive science by using a range of data-driven, bottom-up methods which do not require the domain-

specific analysis of journals and curricula and which may therefore represent a more generalized approach to addressing the interdisciplinary nature of the field. Though we do not claim to resolve the question of whether cognitive science is fundamentally interdisciplinary or multidisciplinary, we argue that the discussion benefits from the novel measurements we present here, which suggest that collaborations and topics within the field have become increasingly integrated in the last 19 years. Specifically, we address the challenges of defining and measuring interdisciplinarity in cognitive science through a combination of co-authorship network features and topic analysis. We validate our measures using full papers from the Cognitive Science Society proceedings between 2000 and 2019 and abstracts from the Vision Science Society (only abstracts are submitted) over a similar time frame (2001 to 2019). We further show that measures derived from network structure and research topics offer a viable means of studying interdisciplinary collaboration and the movement of the field more broadly by using a combination of structure and topic measures to predict new and persisting collaborations in an out-of-sample dataset of the 2020 CogSci proceedings.

First, the degree to which a community is interdisciplinary or multidisciplinary may in large part be revealed by who collaborates with whom. Scientific collaboration can be represented as an undirected graph, in which nodes correspond to individual authors and edges between nodes indicate whether any two authors co-authored a paper together (Barabási et al., 2002; Newman, 2001, 2004). Co-authorships within a community containing multiple areas of study can range from highly integrated to highly modular, and the structure of the resulting co-authorship network will reflect this spectrum of possibilities.

Second, while the collaboration structure of a community no doubt reveals something about the modularity of interdisciplinary work that occurs within it, the ways in which research interests combine must play a role as well. To better understand how the *content* of collaborations informs the interdisciplinarity of the field, we use a topic model (Griffiths & Steyvers, 2004) to extract high level patterns in cognitive science research over the last 19 years. Topic models have been used in previous research to capture trends in the published work within a discipline,

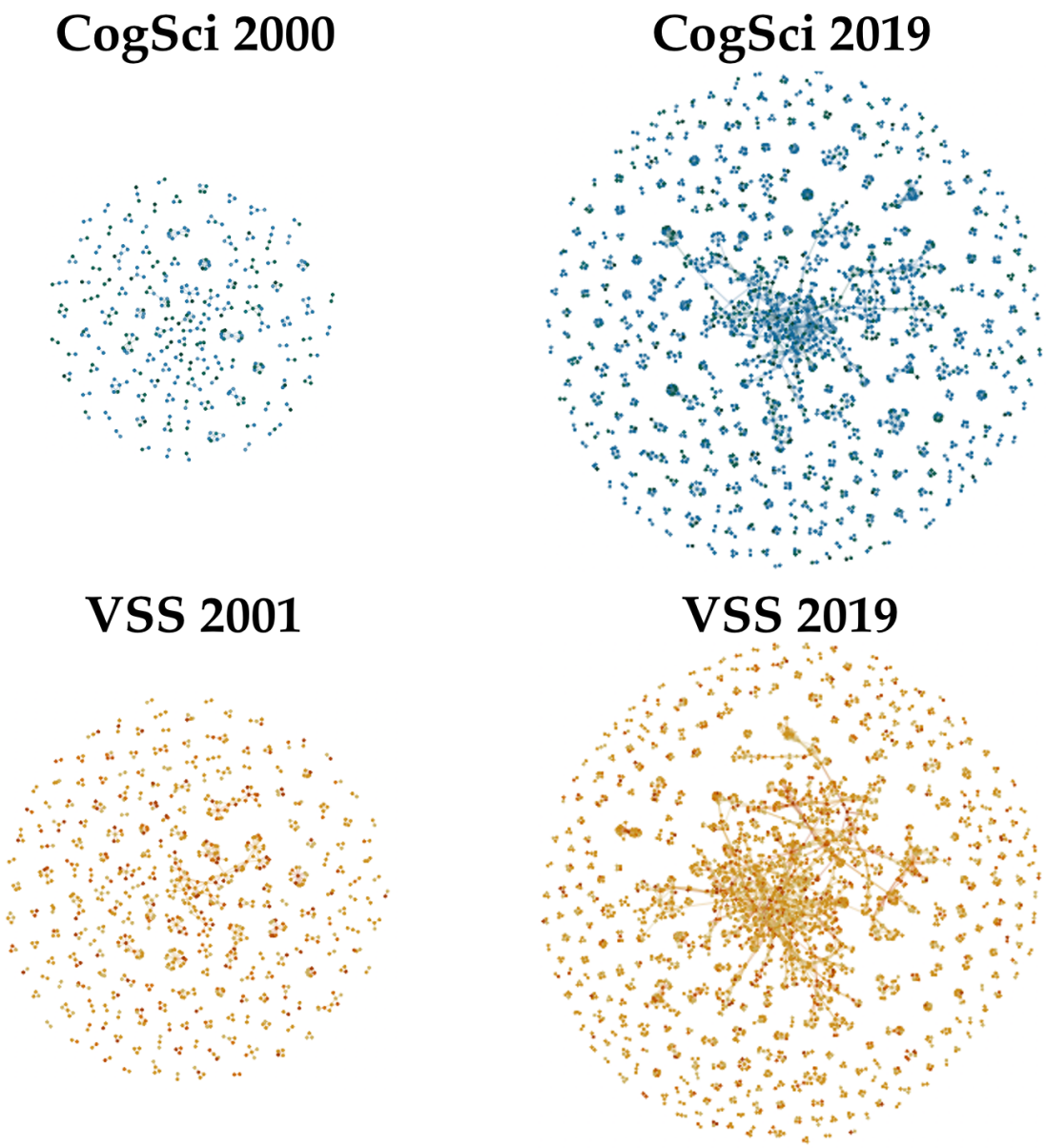


Figure 1.1. The co-authorship network of CogSci in 2000 and 2019 and the network of VSS in 2001 and 2019.

including within cognitive science (Cohen Priva & Austerweil, 2015; Rothe et al., 2018). Studies specifically addressing interdisciplinarity have used topic models to complement pre-defined discipline tagging (Nichols, 2014). In the present work, we apply clustering algorithms to the topics that authors study, addressing the separability of the interests and methods of researchers in the field. More distinct clusters in topic space imply greater division between disciplines.

Finally, in an effort to unify both the structure and the content of collaboration within the

cognitive science community and to illustrate how these variables contribute to our understanding of multi- and interdisciplinary work, we analyze the degree to which structure and topic measures predict future co-authorship. Patterns of interdisciplinary and multidisciplinary collaboration are ultimately revealed in the ways authors form new collaborations and maintain existing ones; thus, the value of topic and network analysis measures for characterizing collaboration in a field can be measured in part by how well they predict novel and continued collaboration. Drawing on our earlier analyses of topic space and co-authorship structure, we assess the role that prior collaboration and topic similarity play in determining whether two authors will collaborate, using the data from the last 19 years to fit a model which we test with out-of-sample data from papers presented at the 2020 meeting of the Cognitive Science Society.

Together, our analyses address (1) interconnectedness in the co-authorship network structure, (2) clusters in the author topic space, and (3) how collaborations arise from a combination of co-authorship network and topic space measures. Not only do these metrics quantitatively illustrate how authorship within cognitive science has changed over time, but we also believe these measures may provide a meaningful contribution to the multidisciplinary-interdisciplinary debate across science[1].

1.2 Data

We retrieved 11,553 full text PDFs (with 12,203 unique authors) from the published *Proceedings of the Annual Meeting of the Cognitive Science Society* from 2000 to 2019[2]. This data is primarily full text conference proceedings papers but also includes submitted abstracts. In addition, we retrieved 22,504 Vision Science Society Annual Meeting abstracts (with 23,842 unique authors) published in the *Journal of Vision* from 2001 to 2019[3]. Both data sets were

[1]All code used in this analysis can be found at: `https://github.com/isabelladestefano/formalizing_interdisciplinary_collaboration`

processed to extract unique authors, publication year, and the full text of each paper or abstract.

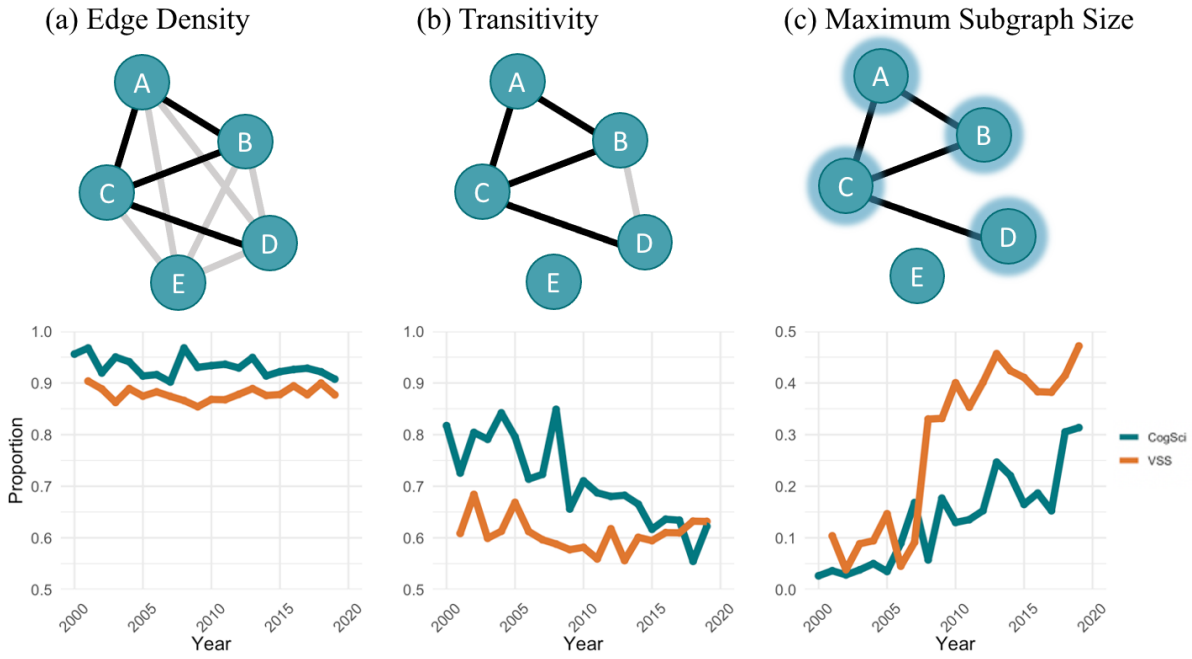


Figure 1.2. Co-authorship network measures and their change over time. For each of the network representations, nodes are connected by the black edges. (a) Edge density, or the proportion of edges in the graph to the theoretical maximum given the number of papers and authors per paper. (b) Transitivity, or the proportion of authors whose co-authors also publish together. (c) The maximum subgraph size, or how many authors are in the largest island relative to the full graph.

1.3 Co-Authorship Network

Using the publication data collected from CogSci and VSS proceedings, we generated a co-authorship network for each year of the conferences with nodes representing authors and edges representing co-authored publications by pairs of authors in that year's proceedings. The graphs were unweighted, i.e., edges represented whether two authors published together *at all* in a given year. We analyze three graph-theoretical measures which, when applied to the collaboration networks, provide insight into the level of interdisciplinarity within these conference communities: edge density, transitivity, and maximum subgraph size.

[2](a) 2000-2014 papers are hosted at `https://escholarship.org/uc/cognitivesciencesociety/`, retrieved 9 December 2018; (b) 2010-2019 papers are hosted at `https://cogsci.mindmodeling.org/`, retrieved 9 December 2018 and CogSci 2019 retrieved 3 December 2019. Processed paper data is hosted at: `https://osf.io/qwzgd/`.

[3]All abstracts hosted at `https://jov.arvojournals.org/`, retrieved 6-8 January 2020

1.3.1 Edge Density

Edge density refers to the proportion of edges within the network relative to the theoretical maximum. Here, the theoretical maximum is determined by the number of edges possible given the total number of publications in that year. For every paper, there exists a fully connected subgraph of the paper's authors with $\frac{n(n-1)}{2}$ edges, where n is the number of authors on that paper. Thus, the full set of N papers, and their associated number of co-authors, sets a theoretical maximum number of edges at $\sum_{i=1}^{N} \frac{n_i(n_i-1)}{2}$. We define edge density for a given year by normalizing the observed number of edges by this theoretical maximum (Equation 1.1).

$$edge\ density = \frac{|E(G)|}{\sum_{i=1}^{N} \frac{n_i(n_i-1)}{2}} \tag{1.1}$$

where $|E(G)|$ is the total number of edges in the co-authorship network G for that year, N is the total number of papers published in that year, and n_i is the number of authors on any given paper i. Our edge density metric measures the degree of repeated collaboration between any two authors, as a proportion of the amount of possible collaboration: a higher edge density indicates a higher rate of unique co-authorships. In an interdisciplinary community, we expect a *higher edge density*, indicating that authors tend to publish with a broad set of collaborators.

The edge density metric is shown in Figure 1.2. The edge density for both CogSci and VSS appears relatively stable over the range considered. Critically, we note that the edge density for VSS is significantly lower than CogSci ($\hat{\beta}$ = -0.05, $p < 0.001$) and, perhaps more importantly, the CogSci edge density measure is relatively close to the theoretical maximum for this measure. This suggests that on average, CogSci authors publish with many unique authors.

1.3.2 Transitivity

Transitivity measures the probability of a node's adjacent nodes also being connected by an edge, i.e., closed triads. Also referred to as the clustering coefficient, transitivity approximates the commonality of local clustering in the graph, such that higher transitivity indicates more

clustering. Thus, we would expect an interdisciplinary community to have *lower transitivity*—authors publish with authors across group boundaries.

The transitivity for CogSci appears to decrease over time whereas the transitivity of VSS remains low over the range considered. Indeed the slope of a regression against year is significantly negative ($\hat{\beta}$ = -0.012, $p < 0.001$), suggesting that the transitivity of the CogSci network is decreasing meaningfully. This could be influenced by a number of factors, including the possibility that authors have published more papers in the proceedings over time. Nonetheless, the decreasing transitivity suggests that collaborations are often between a more diverse set of individuals: that is, CogSci has become less "clique-y".

1.3.3 Maximum Subgraph

The *size of the maximum subgraph* specifies the proportion of nodes in the graph that are connected to the largest island. A network with a large island relative to the overall size of the graph indicates that many authors are connected to many other authors through their co-authors' and their co-authors' co-authors' collaborations. We would expect an interdisciplinary community to have a *large maximum subgraph*, reflecting the tendency of a large subset of the field to be connected in the same collaboration network.

Across both VSS and CogSci, the maximum subgraph appears to grow over the analyzed time period. Broadly, this suggests that the network of authors within the CogSci community has become increasingly interconnected: the positive slope of this increase in the CogSci data is significant ($\hat{\beta}$ = 0.014, $p < 0.001$).

1.4 Topic Space

To extract the research topics studied by the cognitive science community, we used the `stm` package in R (Roberts et al., 2014) to fit a topic model to the full text of the papers from the CogSci and VSS proceedings. `stm` provides functions for cleaning the data by removing punctuation, stopwords, and numbers, then lemmatizing the remaining text. Finally, we fit a

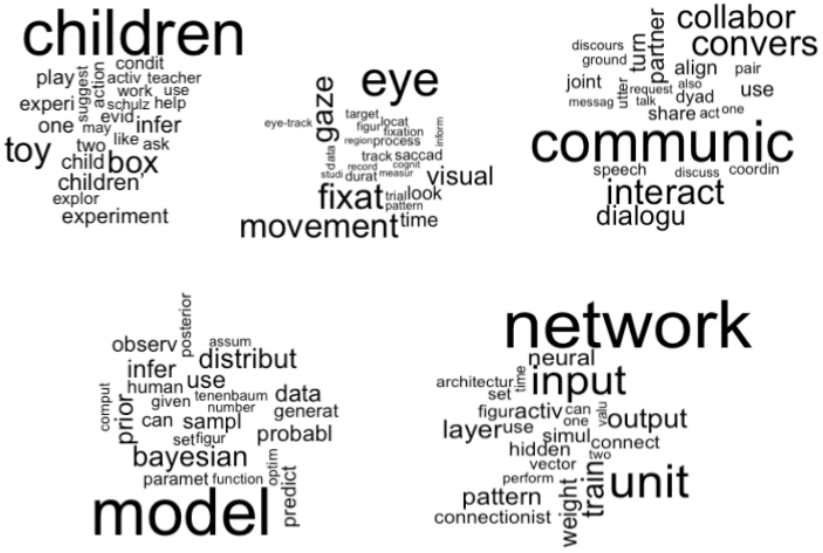


Figure 1.3. Word clouds of frequent lemmas from selected topics. These examples illustrate the level of granularity that the topic model is able to extract from the CogSci texts with 100 topics.

Latent Dirichlet Allocation (LDA) topic model to the full text documents (Blei et al., 2003; Griffiths & Steyvers, 2004). In the model fitting process we specified 100 topics, which yielded niche yet enduring topics and methods, e.g., theory formation (Gopnik & Sobel, 2000), rational analysis (Chater & Oaksford, 1999), and connectionism (Rumelhart & McClelland, 1986). See Figure 1.3 for several examples of high probability lemmas belonging to particular topics fit by the model. The topic model estimates a distribution over the 100 topics for each paper (or abstract); author locations in topic space were computed to be the overall distribution of their topics across all papers they had published in a given year. To alleviate unusually high spikes within topic distributions resulting from authors that publish only one paper, we smoothed the distributions by regularizing individual authors' topic distributions in a given year to the overall topic distribution for each year.

To understand how *integrated* the topics were year over year, we first applied multidimensional scaling (MDS) to the authors' distributions across the 100 topics to reduce the space to two dimensions, which is easier to visualize. We computed clusters on the scaled topic space of authors via k-means clustering (we used $k = 5$ which seemed to balance resolution of salient

Figure 1.4. K-means cluster analysis ($k = 5$) on topic space of authors, mapped onto 2 dimensions via MDS. The cluster maps show the clustering of topics studied by authors in the earliest (left) and most recent year (right) for both CogSci (blue) and VSS (orange). The line graph shows the ratio of within- to between-cluster sums of squares for each year. CogSci is becoming less clustered over time.

clusters and consistency across years). If authors are more clustered in topic space, that reflects less connectivity between disciplines and suggests a multidisciplinary community. To measure the separability of clustering across years, we computed the ratio of the within-cluster sum of squares to the between-cluster sum of squares based on the k-means centroids. A higher ratio reflects greater dispersion within clusters compared to between clusters, indicating that the clusters are not very separated—in other words, that authors are less siloed in disciplinary enclaves, as would be the case in a more interdisciplinary field.

The central plot in Figure 1.4 shows the ratio of the total within-cluster sum of squares to the between-cluster sum of squares for CogSci and VSS between 2000 and 2019. While VSS appears relatively stable (a regression on the data during this range is in fact negative: $\hat{\beta} = -0.006$, $p = 0.004$), the CogSci data has increased dramatically during this time ($\hat{\beta} = 0.014$, $p < 0.001$). Our results suggest that clusters in topic space have become less separable over time. The left and right sides of Figure 1.4 are the author clusters for the earliest and most recent years of the CogSci and VSS data sets. The increase in topic overlap (decreased separability) in the set of CogSci authors is apparent in the two plots while topic consolidation in VSS does appear more nominal.

1.5 Combining Topic Space and Network Structure

In the previous sections, we argue that structural measures of collaboration and general trends in topic space are both useful in trying to quantify interdisciplinarity. However, interdisciplinarity is not only about community structure and topic distributions alone, but about the distribution of topics studied *within* the co-authorship structure. Here, we ask how topic similarity and prior collaboration structure combine to contribute to new and persisting collaborations between authors. In this way, we test the putative role that structural and topic space variables play in determining the overall collaboration landscape of the field. Concretely, we frame the contribution of topic similarity and co-authorship structure in the CogSci network as a link prediction problem: how do these variables contribute to the likelihood that two authors will publish a paper together in a given year? To the degree that both variables can be combined to produce high-fidelity predictions of new and ongoing collaborations, this represents a novel means of synthesizing research content and structure to predict the overall movement of the field. In addition, this validates the use of measures derived from topic space and network analysis to better describe the level of interdisciplinary work that authors are engaged in.

The first measure we use to predict co-authorship is the topic similarity between potential collaborators. Our earlier measures of separability in topic space suggest that author research topics offer insight into the level of interdisciplinary collaboration in cognitive science. Building on this insight, we investigate how the relative position of individual authors in topic space effects their probability of forming a new collaboration or maintaining an existing one. We measured similarity in topic space between two authors in a given year using their cosine similarity: the cosine of the angle θ between two authors' 100-topic vectors fitted by the topic model.

$$cos(\theta) = \frac{a \cdot b}{||a||||b||} \tag{1.2}$$

The cosine similarity between each pair of authors was log-transformed to eliminate skewedness.

We fit a logistic regression to the co-authorships during each year with the topic similarity between authors from the previous year and whether the authors published together the previous year as predictors. We find that prior collaboration, topic similarity, and their interaction, are all significant predictors of future collaborations between authors (*prior collaboration:* $\hat{\beta} = 4.592$, *topic similarity:* $\hat{\beta} = 2.248$, *interaction:* $\hat{\beta} = -3.862$, all ps < 0.001). Though the magnitude of the coefficients themselves might offer insight into the nature of collaboration in cognitive science, we refrain from interpreting the $\hat{\beta}$ values for the simple reason that confirming the independence of these predictors remains a challenge. Insofar as particular patterns of collaboration in the network are associated with corresponding patterns of topic similarity, the magnitudes assigned to these coefficients may not reflect the magnitude of their predictive power.

Instead, to ensure the strength of all the predictors in our model, we compare the full model described above—which predicts new collaborations on the basis of prior publication, topic similarity, and their interaction—to lesioned models with only prior publication and only main effects. Alignment of research topics across collaborators should be inevitable to some degree, assuming coherent and stable research interests. So, it is perhaps unsurprising that collaboration is strongly predicted by both prior collaboration and topic similarity. However, it is less clear whether both variables are necessary. Put another way, does topic similarity and its interaction with prior publication predict collaborations above and beyond having previously collaborated? The full model outperformed both lesioned models (topic similarity: $deviance = 1150$, $p < 0.001$; interaction: $deviance = 518$, $p < 0.001$), suggesting that topic similarity and the interaction between topic similarity and prior publication improve predictions of novel collaborations.

1.5.1 Predicting 2020 Collaborations

Using the regression with prior publication, topic similarity, and their interaction as predictors and training data from CogSci 2000 to 2019, we generate predictions about who co-authors together in CogSci 2020. A subset of these predictions are shown in Figure 1.5a. To evaluate the model's effectiveness, we compare the model's predictions to holdout data: the full

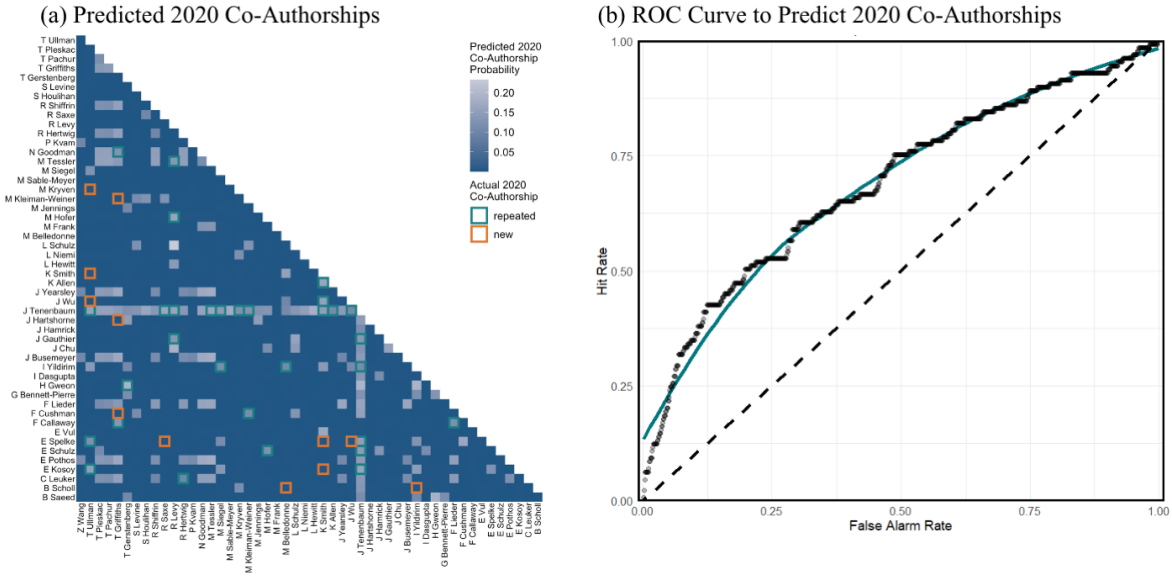


Figure 1.5. Evaluating the co-authorship predictions from combining topic space and network structure. a) Prediction of co-authorships in 2020 for the 50 most eigencentral authors of 2019. The lighter the tile, the more likely our model predicts two authors will publish together. Highlighted tiles were co-authorships that indeed occurred in 2020, where tiles highlighted in teal were repeated collaborations from 2019 and tiles highlighted in orange were new collaborations. b) ROC curve created by using different thresholds on the probability of new publication to make binary predictions. We evaluate only cases where authors did not publish together in the previous year. The dotted line shows where an ROC curve would fall for a model making predictions at chance.

set of collaborations from CogSci 2020 (879 papers, 1929 unique authors). Figure 1.5b shows a Receiver Operator Characteristic (ROC) curve (Swets, 1988) for the model's predictions. A model that predicts co-authorships might attain reasonably high accuracy simply by assuming that authors who previously collaborated together will do so again. For a more stringent test, we consider *new collaborations only* (when there was no prior collaboration the previous year): a prediction which is made based on only the authors' topic similarity. We use the area under the curve (AUC) to evaluate how well our model predicted new publications; an AUC of 0.5 indicates chance performance and an AUC of 1 indicates perfect classification accuracy. We found our model had an $AUC = 0.689$, which indicates our model is well above chance when making predictions about new collaborations. At the optimal threshold, the model's predictions have specificity (i.e., true negatives) of 0.802 and sensitivity (i.e., true positives) of 0.504. If

we instead evaluate all co-authorships, including authors with and without prior collaboration, we obtain an $AUC = 0.869$, indicating that stability of collaboration networks plays an outsized role in publications. The ability to predict new collaborations from out-of-sample data based purely on topic similarity, and to predict all collaborations using a combination of topic similarity and prior collaboration, suggests that variables related to both authorship network structure and research topic space play a role in the ways collaborations form and persist. This bolsters the claim that questions related to interdisciplinary and multidisciplinary research, which are tied to collaborations both new and old in a given field, can be addressed using data-rich, bottom up measures derived from co-authorship patterns and topics.

1.6 Discussion

It has been argued that science is becoming more interdisciplinary across a broad range of research areas (Porter & Rafols, 2009). However, a recent debate in the cognitive science community raises questions about whether the diverse fields that contribute to cognitive science pursue integrated research or are better described as multidisciplinary (Gray, 2019; Núñez et al., 2019). We argue that this discussion—and further investigations into the interdisciplinary nature of research more broadly—is strengthened by the use of formal, bottom-up measures of the collaboration structure and content within the field. Using the full text and author data from 19 years of published proceedings of the Cognitive Science Society, we analyze the evolution of the co-authorship network and assess changes in topic space year over year. Furthermore, we examine the distribution of topics within the co-authorship structure by querying how authors select their collaborators based on their interests and prior collaborations. Since these methods are novel in their application, we further validate their use by comparing the CogSci results to the full set of abstracts published in the Vision Science Society over a similar time period.

Our bottom-up approach yields converging support for the claim that cognitive science researchers have become more integrated over the past two decades. First, the co-authorship net-

work shows that researchers published in the CogSci proceedings have become (structurally) less clustered and more interconnected, as evidenced by the decreasing transitivity of co-authorships and increasing maximum subgraph size. Second, co-authorship edge density, though more stable over time, is consistently higher for CogSci than VSS, suggesting that CogSci authors tend to publish with more unique authors. Third, beyond the structure of collaboration networks in CogSci, we find that the clustering of authors by topic within the CogSci proceedings has become less separable over time. This suggests that distinctions among disciplines may be shrinking. Finally, by combining co-authorship network and topic information, we find that prior collaboration and topic similarity are both significant predictors of collaboration in subsequent years; the significant interaction between them suggests that this is not a simple additive relationship. These variables allow us to predict new collaborations in out-of-sample data from CogSci 2020. Critically, this validates the use of measures derived from both the co-authorship network and the research topic space to characterize interdisciplinary collaboration. More broadly, it suggests that the combination of topic modeling and network analysis provide a window into the ongoing developments in a scientific field from one year to the next.

The use of topic modeling, characteristics of the co-authorship network, and the combination of the two offers a novel set of measures for understanding interdisciplinarity in a given field. The strength of these measures, apart from their formality, is the degree to which they are sensitive to the data in the research itself. Rather than pre-specifying the unique disciplines or fields within the community, we let graph clusters and topic separability speak to the connectedness of the research being done. This may allow for broader application across a range of other fields.

Critically however, the measures we outline here are only part of the larger discussion about whether cognitive scientists are conducting interdisciplinary research. Key to understanding the progression of research in a field over time is not just how interconnected authors become or how creatively existing topics are combined, but how the reach of the network and the topics themselves evolve. Intuitively, interdisciplinarity is about both the *integration* of authors and

topics over time, as well as maintaining or even increasing their *diversity* (e.g., Feng & Kirkley, 2020). A field that becomes more interconnected by barring certain sub-fields or methodologies can hardly be said to have accomplished the goal of interdisciplinary work. The measures we outline here provide a precise and nuanced picture of the integration of authors and topics in cognitive science, but fall short of allowing us to draw strong inferences about the diversity of topics and disciplines represented over time. Indeed, the contrast between the current results and those of works like Núñez et al. (2019) may be in large part attributable to this distinction. Future work should explore ways that the data-driven approach we outline here might be expanded to reflect the goals of broad affiliation and diverse interests that interdisciplinary fields aspire to. The present results provide a step in this direction by showing how the tools of network analysis and machine learning can inform questions about the ways in which collaborations and research topics reflect meaningful integration.

1.7 Acknowledgments

We thank Rafael Núñez, Carson Miller Rigoli, Michael Allen, and Richard Gao for helpful discussion. We also thank Jamal Williams and Hayden Schill for their assistance with an earlier version of the analyses presented here. Finally, we thank the CogSci 2020 organizers for graciously sharing an early list of the authors published in the 2020 proceedings. This material is based upon work supported by a UCSD Research Grant to ID, the NSF Graduate Research Fellowship under Grant No. DGE-1650112 to LAO and UCSD Research Grant No. RG095178 to EB.

Chapter 1 is a reprint of material as it appears in DeStefano, I., Oey, L. A., Brockbank, E., & Vul, E. (2021). Integration by parts: Collaboration and topic structure in the CogSci community. *Topics in Cognitive Science*, *13*(2), 399-413. The dissertation author was a primary investigator and co-first author of this paper.

References

Acedo, F. J., Barroso, C., Casanueva, C., & Galaán, J. L. (2006). Co-authorship in management and organizational studies: An empirical and network analysis. *Journal of Management Studies*, *43*(5), 957–983.

Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A*, *311*(3-4), 590–614.

Bender, A. (2019). The value of diversity in cognitive science. *Topics in Cognitive Science*, *11*(4), 853–863.

Bergmann, T., Dale, R., Sattari, N., Heit, E., & Bhat, H. S. (2017). The interdisciplinarity of collaborations in cognitive science. *Cognitive Science*, *41*(5), 1412–1418.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(January), 993–1022.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65.

Cohen Priva, U., & Austerweil, J. L. (2015). Analyzing the history of *Cognition* using topic models. *Cognition*, *135*, 4–9.

Feng, S., & Kirkley, A. (2020). Mixing patterns in interdisciplinary co-authorship networks at multiple scales. *Scientific Reports*, *10*(1), 1–11.

Goldstone, R. L., & Leydesdorff, L. (2006). The import and export of *Cognitive Science*. *Cognitive Science*, *30*(6), 983–993.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *71*(5), 1205–1222.

Gowanlock, M., & Gazan, R. (2013). Assessing researcher interdisciplinarity: A case study of the University of Hawaii NASA Astrobiology Institute. *Scientometrics*, *94*(1), 133–161.

Gray, W. D. (2019). Welcome to Cognitive Science: The once and future multidisciplinary society. *Topics in Cognitive Science*, *11*(4), 838–844.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(1), 5228–5235.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, *101*(suppl. 1), 5200–5205.

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, *100*(3), 741–754.

Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., & Semenuks, A. (2019). What happened to cognitive science? *Nature Human Behaviour*, *3*(8), 782–791.

Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., & Semenuks, A. (2020). For the sciences they are a-changin': A response to commentaries on Núñez et al.'s (2019) "What happened to cognitive science?" *Topics in Cognitive Science*, *12*(3), 790–803.

Parinov, S., & Kogalovsky, M. (2014). Semantic linkages in research information systems as a new data source for scientometric studies. *Scientometrics*, *98*(2), 927–943.

Porter, A. L., Cohen, A. S., Roessner, D., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, *72*(1), 117–147.

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719–745.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, *82*(2), 263–287.

Ravikumar, S., Agrahari, A., & Singh, S. N. (2015). Mapping the intellectual structure of scientometrics: A co-word analysis of the journal Scientometrics (2005–2010). *Scientometrics*, *102*(1), 929–955.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2014). stm: R package for structural topic models. *Journal of Statistical Software*, *10*(2), 1–40.

Rothe, A., Rich, A. S., & Zhi-Wei, L. (2018). Topics and trends in Cognitive Science (2000-2017). In T. T. Rogers, M. Rau, J. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 979–984).

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. MIT Press.

Schunn, C. D., Crowley, K., & Okada, T. (1998). The growth of multidisciplinarity in the Cognitive Science Society. *Cognitive Science*, *22*(1), 107–130.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293.

Wagner, C. S., & Leyesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, *34*(10), 1608–1618.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, *165*, 14–26.

Chapter 2

Designing and detecting lies by reasoning about other agents

How do people detect lies from the content of messages, and design lies that go undetected? Lying requires strategic reasoning about how others think and respond. We propose a unified framework underlying lie design and detection, formalized as recursive social reasoning. Senders design lies by inferring the likelihood the receiver detects potential lies; receivers detect lies by inferring if and how the sender would lie. Under this framework, we can predict the rate and content of lies people produce, and which lies are detected. In Experiment 1, we show that people calibrate the extremeness of their lies and what lies they detect to beliefs about goals and the statistics of the world. In Experiment 2, we present stronger diagnostic evidence for the function of social reasoning in lying: people cater their lies to their audience, even when their audience's beliefs differ from their own. We conclude that recursive and rational social reasoning is a key cognitive process underlying how people communicate in adversarial settings.

Keywords: deception; rational inference; social cognition

Human communication, though generally honest, is riddled with deception. Most theories of effective communication are predicated on the assumption that interlocutors act cooperatively (Grice, 1975; Grice, 1989). However, in police interviews (Mann et al., 2004), online dating (Hancock & Toma, 2009; Toma et al., 2008), scientific reporting (Fanelli, 2009; John et al., 2012), and news stories (Allcot & Gentzkow, 2017; Lazer et al., 2018), people may choose to present false information. We focus on *lying*, defined here as a sender producing a knowingly false message intended to deceive a receiver. This definition of lying encompasses both verbal and non-verbal communication (Zuckerman et al., 1981) and emphasizes the salient communicative role of the receiver in lying—receivers can believe a lie, or not.

We propose that lying and lie detection arise from interactive, adversarial reasoning where interlocutors must consider how the other will act. In such dyadic communication, receivers are not merely passive audiences—rather, they may punish dishonesty (Ohtsubo et al., 2010; Tyler et al., 2006). Therefore, senders, in deciding which lies to tell, are motivated to avoid being caught by the receiver. Similarly, false accusations are detrimental, and receivers want to avoid them when deciding which messages to call out as lies.

The central idea behind this framework is that the interaction between the competing goals of sender and receiver is critical for deception. While some prior theories highlight how dynamic interaction plays out over the course of back-and-forth conversational sparring (Buller & Burgoon, 1996), our framework highlights the role of anticipated interactivity in human lying cognition, even before a lie is uttered, and places theory of mind (ToM), or the ability to reason about others' mental states and goals (Premack & Woodruff, 1978), at the core of deception. The decision to lie (vs. tell the truth) is known to require some basic ToM understanding, for instance, to acknowledge that receivers may not have access to ground truth and could thus, in principle, be deceived (e.g., Ding et al., 2015). However, the extent to which ToM reasoning drives *how* people actually lie and detect lies has been relatively unexplored.

Theory of mind reasoning is computationally expensive, so people may prefer to rely on other cognitive mechanisms, even when at risk for detection. First, lying is cognitively

demanding (Vrij et al., 2006) and incurs longer response times than telling the truth, even without the risk of getting caught (Capraro et al., 2019; Suchotzki et al., 2017, but see Shalvi et al., 2012). If ToM reasoning itself is a non-automatic, effortful process (Apperly et al., 2006; Lin et al., 2010; Phillips et al., 2015), then applying a complex ToM process would further increase the cognitive demand required of lying. Second, people have been shown to be practically at chance when detecting lies (e.g., Bond & DePaulo, 2006). Under a blanket assumption that detectors are simply guessing, liars need not attribute sophisticated reasoning to lie detectors to succeed at duping them. Third, given the scarcity of distinguishing information about others' idiosyncratic beliefs, a heuristic that relies only on the speakers' own beliefs to choose a lie may well be globally optimal.

2.0.1 Lying and Lie Detection in Isolation and in Dyads

A majority of prior work on lying has omitted key elements of dyadic, adversarial communication that might require theory of mind reasoning, instead focusing on lying (e.g., Gerlach et al., 2019; Mazar et al., 2008) and lie detection (e.g., Bond & DePaulo, 2006; Vrij et al., 2019) in isolation. As a consequence, this prior work cannot speak to whether liars and detectors adapt their strategies in light of considerations of what the other will do. Specifically, research on lie detection has directed its focus on surface features of lies, not the informational content, and has thus paid less attention to the importance of designing lies to be believable. Lie production research, in turn, has primarily used scenarios where liars face no risk of being caught; and thus has also overlooked that real-world lies need to be designed to minimize this risk. In short, by studying lying and lie detection in isolation, prior research has not explored how the two jointly constrain the design of lies in a single communicative act.

Studies of *lie detection* have concentrated on detecting lies from superficial cues, rather than the content of the lie. Classic research on lie detection asked whether lies can be identified from content-independent perceptual cues given off by the speaker, like facial expressions (Bruer et al., 2019; DePaulo et al., 2003; Ekman & Friesen, 1969; Ekman et al., 1988) and verbal pauses

(Granhag & Strömwall, 2002; Vrij, 2008). Other research has prominently been concerned with simple perceptual cues of the message, like whether the statement is repeated (Brashier & Marsh, 2020; Dechêne et al., 2010) or its readability (e.g., statements in **high contrast** are judged as truer than those in low contrast; Reber and Schwarz, 1999; Withall and Sagi, 2021). These extensive bodies of literature have established that content-free perceptual cues to deception are weak and unreliable, despite people's tendency to over-rely on them and their persisting meta-cognitive theories about the diagnosticity of these cues (Vrij et al., 2019). Beyond perceptual cues, other work has focused on lie detection from social and contextual cues, including another person's incentive to lie (Bond et al., 2013; Kraut, 1978), or the (low) base-rate of lying, used as a proxy for making truth judgments (Levine, 2014; Street, 2015). This prior research on lie detection has not examined the relationship between the content of the lie, the receiver's prior knowledge of the world, and their beliefs about the sender's cognitive processes.

Meanwhile, studies of *lying* have examined behavior in scenarios with no risk of being caught, including how (un)willing people are to lie or cheat (Abeler et al., 2019; Gerlach et al., 2019; Mazar et al., 2008), how liars respond to incentives (Gneezy et al., 2013; Mazar et al., 2008), and what lies people produce (Fischbacher & Föllmi-Heusi, 2013; Hilbig & Hessler, 2013; Shalvi et al., 2011). This implicit idea that dyadic interaction is not needed to understand lying behavior is made explicit in the self-concept maintenance account, which proposes that people's lies are constrained by their own beliefs, e.g., about their own honesty (Gino et al., 2009) and moral virtue (Mazar et al., 2008). This account was designed specifically to explain why, even in situations without the risk of detection, people seems to avoid producing large lies. This work posits that aversion to lying, and the selection of lies, is guided by heuristics internal to the speaker.

In contrast to this idea, recent work using a dyadic approach targets how people may consider the listener when lying. These accounts adopt game-theoretic approaches to model strategic behavior in adversarial situations (Becker, 1968). Dyadic frameworks have succeeded at explaining systematic human preferences for general deceptive strategies (e.g., Montague et al.,

2011). For example, research in this vein has shown that senders generally prefer to mislead over outright lying, but when receivers are suspicious, senders elect to be uninformative (Franke et al., 2020; Ransom et al., 2019). Game-theoretic approaches also explain indirect speech for soliciting bribes in circumstances when the speaker is uncertain about the audience's disposition (J. J. Lee & Pinker, 2010). And prominently, lying is more prevalent when it may benefit the listener (as in "white lies": Erat and Gneezy, 2012; Gneezy, 2005), and is less prevalent when others are able to verify the ground truth (Gneezy et al., 2018). This work suggests that reasoning about the interlocutor as having a goal or belief at all—i.e., some kind of theory of mind—may play a key role in lying.

While there is a growing body of literature on dyadic frameworks for understanding deception, many of these studies are designed to understand strategies *other than lying*. Several studies have focused on misleading information (i.e., strategically uninformative content) by considering settings where senders are explicitly prevented from lying (Ransom et al., 2019), or are provided no incentives to lie rather than just mislead (Montague et al., 2011; Rogers et al., 2017). In these cases, lies are unnecessary, and so there is no motive to design them well, or use them at all. On the other hand, most studies of lying have used settings where speakers are not punished for being caught in a lie (e.g., Gneezy, 2005, but see Gneezy et al., 2018), thus again removing incentives to lie strategically. The net effect is that existing research on lying has not explored scenarios where speakers are motivated to design lies that are both advantageous to the sender and plausible to the receiver.

Here we propose, and test, an account of lying as a fundamentally dyadic adversarial reasoning problem. We posit that people detect and generate lies in adversarial settings, by selecting counter-strategies tailored to the behavior of the opponent that they predict, all under the assumption that the opponent is a rational, thinking agent with particular goals and knowledge of the world. We formalize this account in recursively coupled, adversarial theory of mind models of the liar and lie detector. We introduce a novel dyadic lying game, allowing us to measure and parametrically manipulate lying and lie detection behavior in an adversarial context. This

experimental context allows us to test whether people lie and detect lies by reasoning about other agents. In Experiment 1, we use this paradigm to test whether senders consider receivers' beliefs and adjust the plausibility of their lies to the statistics of the world; while receivers reason about the senders' goals, and thus rationally adjust which claims they call out as lies. In Experiment 2, we further test whether senders adapt specifically to the statistics of the world they think that receivers believe to be true, even when they know these beliefs to be false; thus testing the central role of theory of mind representations in the strategic design of lies. Altogether, we find that human behavior exhibits the key qualitative patterns of adversarial theory of mind reasoning predicted by our formal model.

2.1 Formalizing Dyadic Reasoning in Lying and Lie Detection

As a first step, we introduce a formal model of dyadic reasoning in lying and lie detection. Formal models allow us to explicitly define our cognitive assumptions and generate behavioral predictions, which we can empirically test. Most importantly, to test whether dyadic reasoning is driving the behavioral predictions, we can compare the predictions of the dyadic reasoning model to those of alternative models that drop the critical theory of mind reasoning assumptions.

To formalize the interactive, adversarial reasoning inherent in lying and lie detection, we develop an ideal observer model inspired by recursive probabilistic inference models of human social cognition and cooperative communication (Frank & Goodman, 2012; Kao et al., 2014; Shafto et al., 2014). In this account, a *sender S* chooses what to say in light of how they believe a *receiver R* will respond, and the receiver decides whether the utterance is a lie based on what they believe a sender would say in different world states. We formalize the lying interaction as follows: the sender observes the true state of the world k, and chooses how to report the state of the world k^* to the receiver. The receiver can either accept k^* as the true state of the world, or challenge the veracity of the report by calling *BS*. Senders are motivated to report an alternative

state of the world k^* that advantages them most while still being believed by others. Thus they are constrained by two conflicting goals: (1) gain—a bigger lie (larger k^*) yields more rewards if accepted, and (2) plausibility—bigger lies are less plausible and more likely to be detected. Meanwhile, the goals of receivers are (1) to successfully detect lies to not be swindled, but (2) to avoid false accusations.

Receivers Detecting Lies

A receiver decides whether to call BS on a reported signal by computing the expected value of making an accusation for the reported signal, and comparing it to the expected value of accepting it as the truth. This calculation relies on combining the receiver's utility for calling BS (would I benefit from calling this message out as a lie?), with the receiver's beliefs about whether a given signal reflects the true state of the world (how likely is this message to be a lie?). This posterior belief arises from what the receiver believes of the sender's likely actions. The receiver must consider both what they believe the sender would report in each world state $P_S(k^* \mid k)$ and the distribution of true states of the world $P(k)$:

$$\mathrm{EV}_R(BS \mid k^*) \propto \sum_k U_R(BS; k^*, k) P_S(k^* \mid k) P(k) \tag{2.1}$$

Thus, choosing whether or not to call BS in response to a given report requires an estimate of how the sender decides what to report.

Senders Designing Lies

A sender decides what to report by calculating the expected value of each possible message based on their rewards and the likelihood that the receiver will call out a reported signal as a lie $P_R(BS \mid k^*)$:

$$\mathrm{EV}_S(k^* \mid k) \propto \sum_{BS} U_S(k^* \mid BS, k) P_R(BS \mid k^*) \tag{2.2}$$

Thus, the sender chooses not only whether to lie, but which lie to tell—potentially more rewarding but more conspicuous—based on their beliefs about how the receiver will respond to each message.

Equations (2.1) and (2.2) compute the expected value of the receiver's and sender's potential decisions, respectively. Both agents are assigned a probability that they will choose their actions by employing a Luce choice rule (Luce, 1959) over their expected values for each potential decision. In this way, the model builds in an assumption that receivers and senders rationally simulate the outcomes of alternative actions when deciding how to act.

Recursive Theory of Mind

If lying is a fundamentally dyadic, theory of mind reasoning problem, senders should lie and receivers should detect lies based on beliefs about their opponent's mental states and how they predict the other agent will make decisions. This means that equations (2.1) and (2.2) feed into each other: a receiver's decision to call BS is a function of their belief about the sender's actions; a sender's decision to lie is a function of their belief about the receiver's actions.

Whether a receiver calls BS and what a sender reports are defined via mutual recursion. Such recursive definitions might yield infinite computational complexity (if they were rolled out to infinite depth). In cooperative communication settings, mutual recursion converges given the concordant goals of the agents (Frank & Goodman, 2012; Schelling, 1960); however in adversarial settings, such recursion often fails to converge, and instead might cycle. We follow a conventional approach to resolve such non-convergent behaviors and follow the cognitive hierarchy model (Camerer et al., 2004) to define the agents as believing in a Poisson distribution over the depth of recursion that their opponent will consider. In other words, the sender may assume that their opponent is sometimes a 0-step receiver (i.e., calls BS randomly), a 1-step receiver (i.e., calls BS assuming the sender thinks the receiver is random), or an n-step receiver. However, rather than committing to a single assumption about the receiver, the sender assumes that the receiver is a weighted combination of all these potential strategies. The player then

reasons one step further, choosing the best action in response to this weighted evaluation of their opponent's likely behavior. The Poisson distribution over opponent recursion depths smooths out cycling behavior in adversarial settings, and yields convergent results (Camerer et al., 2004). It is worth noting that the Poisson rate parameter is usually tuned to yield behavior consistent with humans, but is not independently verified to accurately track the distribution of reasoning depths of ecologically representative opponents. We refer to this strategy as the `Recursive Theory of Mind (ToM)` account of deception.

2.1.1 Alternatives to Dyadic Reasoning in Lying

Lying Heuristics

Many accounts of dishonest behavior do not assume that it arises from a rational consideration of alternatives, but instead is driven by certain inflexible strategies: heuristics. According to these accounts, individuals restrain their lies following simple self-oriented rules (e.g., Abeler et al., 2019; Gerlach et al., 2019; Mazar et al., 2008). For example, the prominent self-concept maintenance account hypothesizes that people lie by satisfying a constraint to preserve a concept of themselves as moral agents (Mazar et al., 2008). Notably, such accounts posit that lies face constraints from the liar's own values, prior beliefs, and knowledge about the true state of the world. These heuristics form a compelling alternative account, especially to help explain why people avoid saying maximal lies even in settings that do not bear a risk of detection. If people avoid large lies in the same manner regardless of whether they are at risk of being detected, one appealing explanation might be that listeners do not play a role in how people design lies after all; rather, it can be explained by individuals' lying heuristics.

We instantiate versions of these verbal theories as parametric models. The `Equal Intrinsic Aversion Heuristic` account posits that everyone shares the same intrinsic aversion to producing overtly large lies that results in people lying by some small amount on top of the truth. The second model assumes that people can be classified as those that exclusively tell the truth and others that lie (Hurkens & Kartik, 2009; Levine, 2019; Serota et al., 2010), again

by some amount on top of the truth (`Unequal Intrinsic Aversion Heuristic`).

These alternative theories make several critically different predictions from the `Recursive ToM` account. Critically, both of these accounts predict that the size of lies should depend only on what the individual believes to be true, which serves as an anchor from which they adjust slightly. Furthermore, individuals in these accounts are about equally tempted to lie regardless of what the truth is. Additionally, these alternative accounts do not predict that lying behavior should change depending on what lies would be plausible to the receiver from base-rate beliefs about the world, nor is the decision to lie driven by payoffs. If receivers *cannot* respond or senders are not concerned about how the audience responds, then senders have no motivation to reason about, or adjust their behavior to the audience's beliefs or goals. It is worth noting that both the `Equal Intrinsic Aversion` and `Unequal Intrinsic Aversion` heuristic accounts are models of lying behavior and make no prediction about lie detecting behavior.

0^{th} Order Theory of Mind

Let us now consider the minimally different account that specifically lesions the theory of mind component of `Recursive ToM`. This account *does* consider payoff gains for larger lies, so it is a rational agent, that decides what to report based on relative expected utility. However, this alternative agent *does not* attribute sophisticated reasoning to the audience, like having beliefs or goals that drive behavior. Instead, senders assume that the best receivers can do is to behave randomly when detecting lies (first-order intentional system; Dennett, 2009). Such a heuristic assumption about the opponent is not unreasonable: after all, previous work finds that people are practically at chance when detecting lies from the majority of liars (Bond & DePaulo, 2006; Levine, 2010), especially without enough useful context (Blair et al., 2010).

Specifically, this sender believes that their opponent will randomly and uniformly call BS without considering the payoff structure or statistics of the world. This model is equivalent to Equation (2.2) if the sender assumes the receiver is wholly a 0^{th} order reasoner. We call this the `0th Order Theory of Mind` model because the sender is merely attributing a primitive

behavioral strategy to their opponent. If the sender believes their opponent behaves randomly, the sender need not adjust their lies to the statistics of the world; they can get away with lies of the same extremeness in any context. Senders under these conditions should lie maximally all the time, when they think that what they say has no bearing on the risk of being detected. Thus both the heuristic and 0^{th} Order ToM models predict that the sender will produce lies in the same manner regardless of the statistics of the world. For predictions of lying behavior on each of these accounts, see Supplemental Materials.

2.1.2 Predictions

If people believe their opponent uses theory of mind to lie and detect lies, then a rational sender ought to choose how and when to lie conditioned on how they assume a rational receiver ought to call BS. Similarly, the rational receiver ought to call BS conditioned on how they assume the sender ought to lie or tell the truth—which will depend on the sender's beliefs about what the receiver will detect. Thus, both agents select their behavior conditioned on what a rational, albeit noisy, utility-seeking opponent would do, which results in a recursive process of reasoning about the other agent's likely actions. The Recursive ToM account generates four key predictions about lying and lie detecting behavior.

First, when the truth is less favorable, people should lie more often. Conversely, when the truth is already favorable, people have less motivation to lie, so they are more likely to tell the truth. Formally, this intuitive prediction arises out of the sender's goal to select actions that optimize their payoffs. Senders that do not consider payoff gains for producing larger lies will lie about equally often regardless of what was the truth, as in the Equal Intrinsic Aversion Heuristic and Unequal Intrinsic Aversion Heuristic accounts.

Second, people should balance payoff gains and plausibility when selecting what lie to send. While a larger lie might produce greater gains, a lie too large and implausible will be readily detected. Furthermore, when there are changes in the world that affect what might seem plausible, people should also adapt their lies. A hallmark of the Recursive ToM model is that it

predicts rational senders should be attuned to prior beliefs and to the statistics of the world (i.e., the base-rate probability of an event or outcome). Of the models we compare, the `Recursive ToM` model uniquely predicts that senders' lies should be sensitive to base-rate information. Meanwhile, the `Equal Intrinsic Aversion Heuristic`, `Unequal Intrinsic Aversion Heuristic`, and 0^{th} `Order ToM` models predict insensitivity to the base-rate.

Third, plausibility is subjective — so, people should cater their lies to what is plausible to the specific audience in mind. What might seem plausible to a gullible audience may be scrutinized by a more knowledgeable audience, so people should hedge their lies accordingly. Showing that people's lies are sensitive to the specific and unique beliefs of their audience would be strong diagnostic evidence for a role of theory of mind in lying.

These predictions up until now have been about the sender's behavior that follows from considering the sender's goals and reasoning about what lies might be detected by the audience. Are these valid assumptions about the audience's reasoning? As such, a fourth prediction is that these audiences are indeed sensitive to plausibility and payoffs. Receivers should robustly adjust their degree of suspicion based on changes to what seems plausible in the world. Alternatively, receivers may simply accept all reports as true, or randomly and uniformly call BS, ignoring goals and the plausibility of reports. Receivers may also be ignorant to just the payoff structure, in which case they should make neutral assumptions about the goals of the players. Instead, the best the receiver could do is to call out reports that are suspicious simply because they are unlikely to occur by chance. This process is akin to `Null Hypothesis Significance Testing`, in which the receiver has no bias to prefer lies of a certain direction. Lastly, receivers that ignore plausibility should not adjust how they call BS when the base-rate probability of an outcome changes.

2.2 Experiment 1

The core predictions of a rational, theory of mind based model of lying and lie detection are therefore: (1) People lie more when the truth is less favorable for them, (2) People craft lies by considering their plausibility and reward, and (3) People identify claims as lies based on their plausibility and payoffs. In Experiment 1, we test these predictions in a novel, dyadic lying game (Fig. 2.1) where people take turns reporting the number of red marbles drawn from a box, and classifying such reports as truths or lies. To test the core predictions of the `Recursive ToM` model against the predictions of alternative accounts, we manipulate the base-rate of red marbles in the box, and the payoffs associated with marbles of each color. Critically, we set up the incentive structure for senders and receivers so that senders are motivated to tell the biggest lie they can get away with, and receivers are motivated to call out lies, while avoiding false accusations. While these incentives may not reflect all real world lying situations, they capture common tradeoffs, and allow us to isolate the role of theory of mind in lying and lie-detection behavior.

2.2.1 Methods

Participants

A total of 228 participants were recruited from the undergraduate population at the University of California, San Diego. Two participants were excluded for producing out-of-bounds responses. Additionally, 14 participants were excluded for failing to meet the attention check criterion, which entailed achieving (within an absolute error of one) at least 75% (9 out of 12) numeric-response comprehension questions distributed throughout the task. After exclusion, 212 participants were included in the final data set. Participants were randomly assigned approximately evenly across the conditions (see Procedure). The study was conducted online, and participants were rewarded class credit for their participation. Informed consent was obtained from all participants, and all studies were approved by the university's Institutional

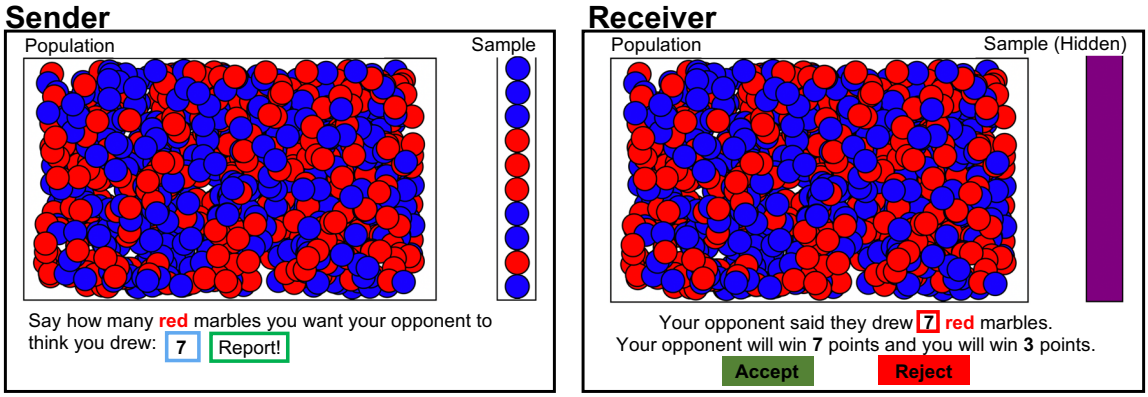


Figure 2.1. Experiment 1 dyadic lying game. The sender and receiver (one of whom is an AI; roles alternate across trials) both see the population of red and blue marbles (in the box; here, 50% red), but only the sender sees the true sample of 10 marbles (in the tube). (Left) Senders report the number of red marbles they sampled; they can tell the truth or lie by reporting something false. In this example, the sender gets points for red, while the receiver gets points for blue. The sender lies by reporting 7 red marbles, when in fact the sender actually sampled 4. (Right) Receivers accept or reject the reported number (i.e., call BS). If the receiver accepts, the sender gets 7 points for the reported red marbles, and the receiver gets 3 points for the reported blue marbles. If the receiver rejects, the sender gets caught in a lie and is penalized, while the receiver is rewarded.

Review Board.

Procedure

Participants played in a dyadic lying game that rewarded participants for strategic production and detection of lies. In each round of the game, both players saw a box of red and blue marbles which had some base-rate probability of sampling a red marble. The sender randomly sampled 10 marbles, of which k were red and the remainder were blue. When prompted about how many red marbles they sampled, the sender reported a number k^* which could be true or false. The receiver then saw how many red marbles the sender reported (with no knowledge of the true number) and could either accept this report or reject it as a lie.

Crucially, if the sender's report was accepted, the sender gained points for the number of red marbles reported k^* and the receiver gained points corresponding to the blue marbles reported $10 - k^*$ (in the condition where senders get points for red). So senders were motivated to lie and report an inflated value. However, if the report was rejected and the sender was indeed

lying, the sender would lose points and the receiver would gain points. If the receiver rejected the report but the sender was in fact telling the truth, the receiver would face a penalty for making a false accusation, while the sender would receive points as they reported (Fig. 2.2 for full payoff structure).

In a 3×2 design, we manipulated (between-subjects) the base-rate probability of drawing a red marble (20%, 50%, 80%) and which color the sender got points for (payoff condition, *red* or *blue*) (Fig. 2.2). Across the payoff conditions, the mapping of points was reversed. When the sender got points for blue marbles (and the receiver for red marbles), the sender still reported the number of red marbles, so the sender was motivated to report *deflated* values (i.e., fewer red marbles corresponds to more blue marbles).

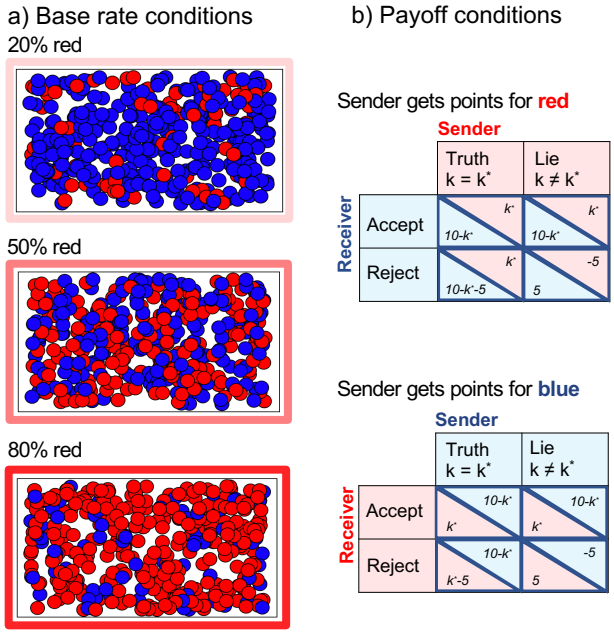


Figure 2.2. Design of experiment 1 conditions. (a) Three base-rate conditions: the probability of sampling red marbles is 20%, 50%, or 80%. (b) Two payoff conditions: the sender gets points for *red* or *blue* marbles. Values in each triangular cell of the payoff table shows the points rewarded to each player (sender: top right triangle; receiver: bottom left triangle). Senders always reported red marbles k^*. Thus, when the sender gets points for red, the sender is motivated to report a higher number (more red), and when they get points for blue, a lower number (fewer red, therefore more blue). If the receiver catches the sender in a lie, the receiver is rewarded 5 points and the sender loses 5 points; if the receiver makes a false accusation, the receiver faces a -5 penalty atop what they would have received.

Participants played for 100 trials, switching roles between every trial, against who they were led to believe was another person but was in fact a computer. Participants were given the goal to win by the highest point difference possible. To discourage participants from learning from the computer's behavior, participants were not given feedback about their opponents' choice (i.e., the true number drawn, whether they lied or told the truth, accepted or rejected report), after the initial practice trials. However, to motivate participants to pay attention, they were given updates on both players' cumulative points after every fifth trial. We expected that feedback only about cumulative points every fifth trial would not allow participants to learn or change strategy over time within the task. In line with this, we find that participants showed no performance improvement as the receiver, and only slight improvement as the sender—amounting to a $+0.7$ score improvement over 100 trials, and that could be attributed to an increased familiarity with the task (see Supplemental Materials for learning analysis).

Additionally, participants were intermittently asked trial-related attention check questions about how many red marbles they drew (if they just played as the sender) or the other player reported (if they just played as the receiver). Participants entered in a textbox their numeric response, and their possible responses were restricted to being between 0 and 10. The questions (12 in total) were randomly distributed after trials throughout the experiment (both practice and test trials).

2.2.2 Results

Senders' Lying Behavior

The behavior of senders can be divided into (a) their rate of lying (as opposed to truth-telling) and (b) the lie they told when choosing to lie. As we cannot pinpoint participants' underlying intentions, here we included lies as any reported value that was false, regardless of its intention. Reports grouped into this category may have been intentional lies designed to advance the player in the game, accidental false reports, etc. We compute the rate of lying (a) as the proportion of false reports to all reports. The lie told (b) is the report itself, conditioned on the

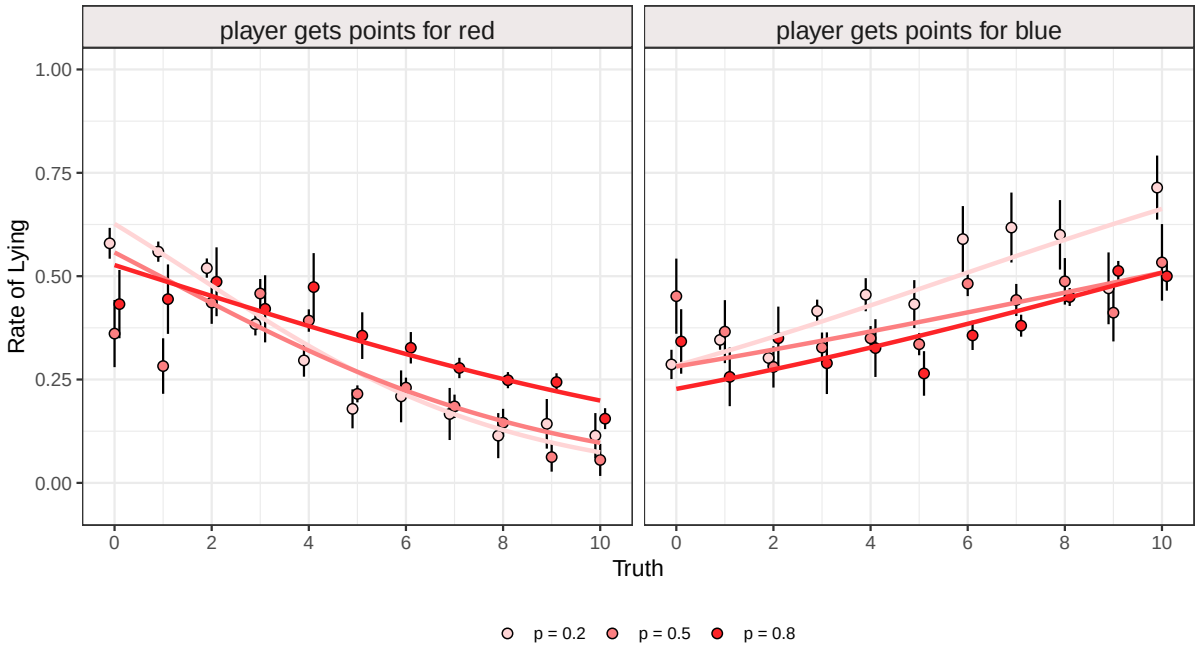


Figure 2.3. The rate of lying given the true sample for each condition in Experiment 1. Each point represents the rate of lying at a given truth value—the true number of red marbles sampled—by condition, and the error bars represent the standard errors of the mean. When senders get points for red, the rate of lying decreases as the truth increases; and vice versa for when senders get points for blue. These results show that people lie more frequently when the truth is less favorable.

report being false (and thus on the true number of red marbles sampled).

Senders lie more when the truth is less favorable to them. If senders lie more when reality is less favorable to them, we would expect the rate of lying to change as a function of how many red marbles they actually saw, such that the rate of lying increases with the number of red marbles seen when senders are rewarded for blue marbles, and to decrease with the number of red marbles seen when they are rewarded for red marbles.

We used the true drawn k, the payoff condition, and their interaction as predictors for the rate of lying in a mixed-effect logistic regression, with subject as a random intercept (Fig. 2.3). The payoff structure was treated as a sum coded factor. Critically, when senders got points for red, there was a negative slope of -0.28 ($SEM = 0.02$, $z = -17.14$, $p < 0.0001$), showing that people decreased their rate of lying when the true k was larger. In contrast, when senders got

points for blue, there was a positive slope of +0.15 ($SEM = 0.02$, $z = 9.36$, $p < 0.0001$), so people increased their lying rate with larger k. Together, these results showed that people, guided by their payoffs, lie more when the truth is less favorable to them.

Senders lie by considering the plausibility and payoff of lies. The Recursive ToM model predicts that senders calibrate the extremeness of their lies to ambient base-rates (the probability of that outcome in the world). If the prevalence of red marbles in the box decreases, the receiver should be more suspicious about higher reported values, and therefore the sender should hedge by reporting fewer red marbles when they lie. Thus, under the `Recursive ToM` model, we would predict on average reported lies would become greater as the base-rate for drawing red marbles increases. In contrast, the `Equal Intrinsic Aversion` and `Unequal Intrinsic Aversion` heuristic models, and the 0^{th} `Order ToM` model, all predict no change in behavior as a function of the base-rate.

To test these predictions, we examined how the relationship between the true drawn k and reported lies (i.e., reported red marbles k^* when they differed from the truth) varied across the base-rate and payoff conditions (Fig. 2.4). We fit a linear regression to the number of marbles falsely reported (i.e., k^* when $k^* \neq k$) with the predictors of the true value of k, the base-rate, the payoff structure, and the full interaction between these three factors. Subject was included as a random intercept. To facilitate comparisons across conditions, the true values of k were centered on 5 so that the models' intercepts correspond to the lies told when 5 marbles were truly drawn. Thus, changes in the intercept reflect changes in which lies are likely to be told in response to seeing 5 red marbles actually drawn.

First, we examined the general relationship between what lies the sender reported and what they actually drew (as the sole predictor in a fixed effect model). As expected, people's falsely reported numbers were larger when they drew more marbles in reality ($\hat{\beta} = 0.33$, $t(3865) = 27.17$, $p < 0.0001$, $r = 0.40$)[1]. This was true regardless of whether someone is motivated to lie by over-reporting (as in the red payoff condition; $\hat{\beta} = 0.41$, $t(1779) = 22.42$, $p < 0.0001$, $r = 0.47$) or under-reporting (blue payoff; $\hat{\beta} = 0.33$, $t(2084) = 19.80$, $p < 0.0001$,

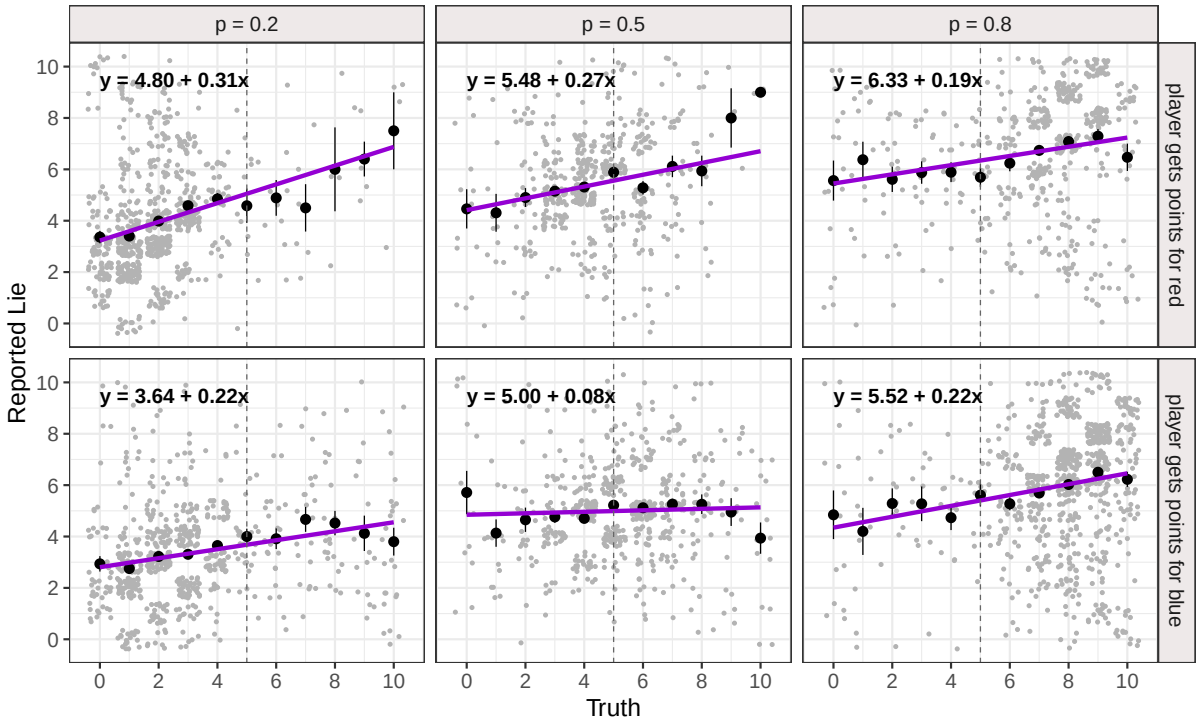


Figure 2.4. The distribution of participants' lies from Experiment 1 across each condition. Each gray point was a false reported value. A linear mixed effect model was fit to each condition, with intercepts centered at $Truth = 5$. Intercepts increased across higher base-rate conditions, and there was a general shift across payoff conditions (top row vs. bottom row). These intermediary results allowed us to interpret differences in lies told across conditions and compare them to the model predictions in Fig. 2.5.

$r = 0.40$).

To address our main question, we next analyzed whether the base-rate condition influenced people's lies. Critically, the sender's reported lie significantly changed across the base-rate conditions ($\chi^2(8) = 87.8$, $p < 0.0001$). When senders got points for red marbles, their lies were

[1]When thinking about the magnitude of people's lies (the distance away from the truth) as a function of the true value, it is important to note that the restricted reporting range of the task means that the magnitude of possible lies is more restricted toward the ends of the range. For example, if the goal of the sender is to over-report how many red marbles they saw, then when fewer red marbles are sampled, there is a greater margin for over-reporting. In this case, people cannot possibly lie by the same magnitude when they see a large number as compared to when they see a small number. In Fig. 2.4, a slope of 1 would indicate a constant difference between truth and lies regardless of how many red marbles were actually drawn. A slope of 1 for these task results would be impossible unless the average lie magnitude was 0—when the truth was 10, speakers cannot possibly lie in the positive direction, since they can only report numbers between 0 and 10. Our results showed a much shallower slope of 0.33, revealing that the magnitude of the lie was smaller for larger truths. However, this behavior may have arisen from the restricted range of the task.

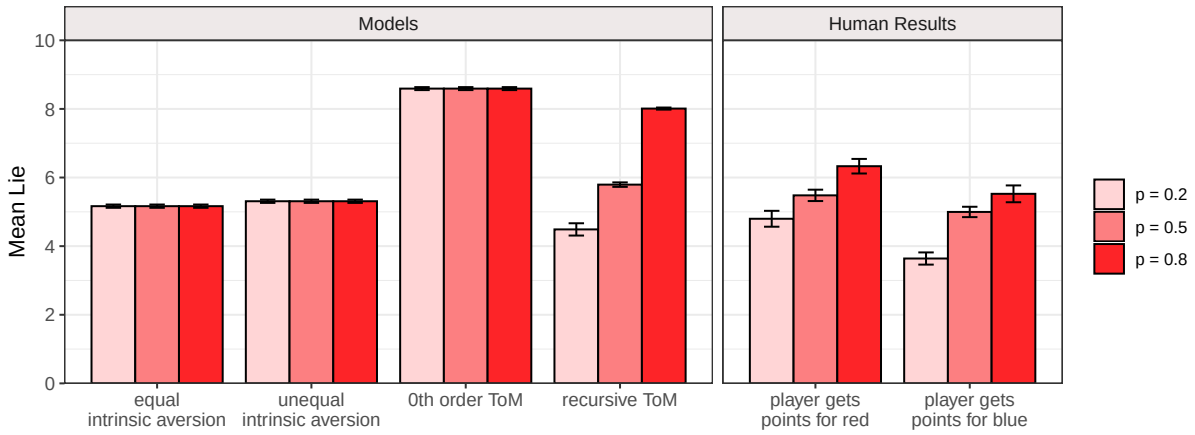


Figure 2.5. The model prediction and human results for the mean lie, computed from the intercept of the linear fit (e.g., from Fig. 2.4). The Recursive ToM model uniquely predicts that the sender should alter their mean lie based on the receiver's base-rate belief. Results from Experiment 1 show that human participants calibrated their lies to the base-rate for sampling red marbles.

highest when the base-rate was 80% (Mean lie = 6.33, SEM = 0.21), intermediary when the base-rate was 50% (Mean lie = 5.48, SEM = 0.17), and lowest when the base-rate was 20% (Mean lie = 4.80, SEM = 0.23). These results are in line with the predictions of the `Recursive ToM` model—that senders calibrate their lies by reasoning about what the receiver may find plausible from base-rate information about the world. The payoff condition (red vs. blue) was also a significant predictor of reported lies ($\chi^2(6) = 32.0,\ p < 0.0001$). In other words, the sender's goal additively shifted the sender's reported lies. This means that in the blue payoff condition the mean lie was also larger for higher base-rate conditions (of red marbles), except with an overall drop in magnitude relative to the red payoff condition. When the sender got points for blue marbles, the mean lie at a base-rate of 80% was 5.52 (SEM = 0.25), 5.00 when the base-rate was 50% (SEM = 0.15), and 3.64 when the base-rate was 20% (SEM = 0.18). Thus both the base-rate and payoff conditions additively affected what lies people reported.

Our evaluation of the sender's lying behavior confirmed the core predictions of the rational, theory of mind based lying model and violated the predictions of the alternative heuristic models (Fig. 2.5). Namely that senders lie more when the truth is less favorable to

them, and they choose lies by considering both their plausibility and the players' payoffs.

Receivers' Lie Detecting Behavior

The `Recursive ToM` model of rational, statistical senders assumes that receivers are themselves rational, statistical agents. Specifically, it assumes that receivers are more likely to identify a claim as a lie if it is less plausible and more consistent with the reward structure. However, the prevailing view is that human lie detection behavior is close to chance (54% accuracy; Bond and DePaulo, 2006; Gladwell, 2019; Levine, 2014). If receivers are at chance, random, or otherwise insensitive to how a claim compares to the relevant statistical information in the world, then reports should not be called out based on simple base-rates. Alternatively, the receiver may have preferences that are not dependent on the relevant goals. In this case, the report least likely to be called out should simply be the most likely number arising from random sampling (e.g., 5 red marbles, when 50% of the marbles in the population are red). This is the prediction of the `Null Hypothesis Significance Testing (null)` account. Within each of these accounts, receivers do not exhibit the sophistication we attribute to the senders' model of the receiver. Alternatively, if the receiver prefers reports of fewer red marbles, or if the receiver knows the sender is motivated to report more red marbles, then reports of more red marbles are more likely to be called out, and the reports that the receiver accepts as true will have fewer red marbles on average. Do real human receivers detect lies in the rational manner we have assumed in our senders' model?

Figure 2.6ab shows the rate at which receivers reject a given reported number of marbles, revealing that receivers are more likely to reject as a lie reports of many red marbles when the base-rate of red marbles is lower (indicating a sensitivity to the plausibility of reports), and when red marbles are rewarded for the sender (indicating a sensitivity to payoffs). To quantify these patterns, we characterized the receiver's behavior in terms of the report that they were most likely to accept (i.e., least likely to call out as a lie) in Figure 2.6c. We estimated this value by taking the maximum-likelihood of a (vertex-form) quadratic logistic regression fitted to the

human receiver data. This allowed us to infer the report for which receivers least called BS for each condition (see Supplementary Materials for more details).

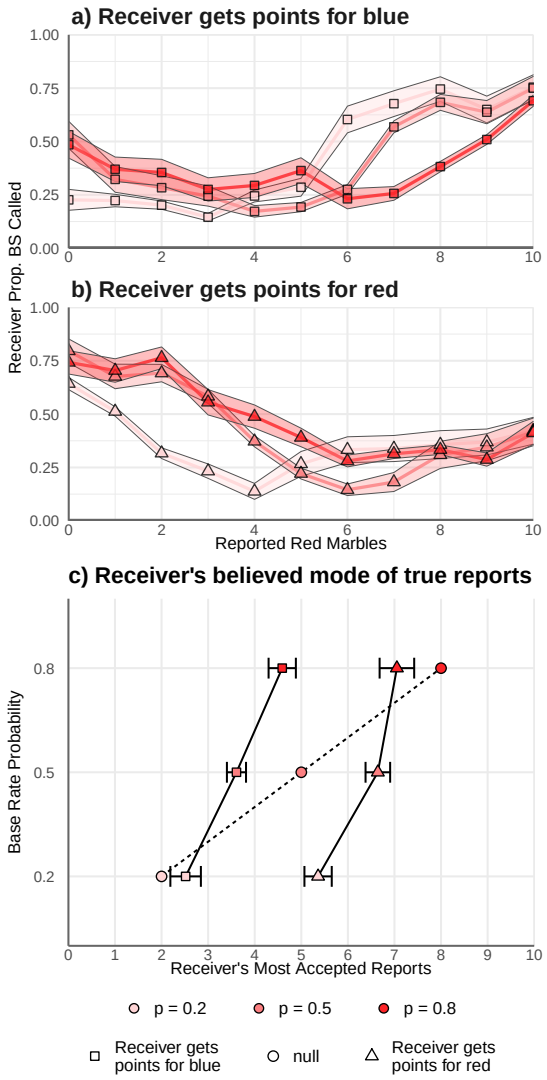


Figure 2.6. Human results for the receiver's rate of calling BS (rejecting the reported value as a lie) based on how many red marbles the sender reported. As predicted by `Recursive ToM`, (a) when receivers got points for blue, receivers more often called out reports of high numbers of red marbles as lies; (b) when they got points for red, the opposite was true. Receivers accounted for the statistics of the world in detecting lies, shown by the shift in BS-calling across base-rate conditions. (c) We summarized the results of (a, b) by estimating which reported value receivers are most likely to accept. This quantity can be interpreted as the implied mode of the believed true reports (x-axis). The error bars represent the 95% confidence interval of the mean. Under the null—receivers are ignorant to the payoff structure—the mode would be equal to $10\times$ the base-rate and would not vary by payoff condition. Instead, receivers' behavior varied systematically with the payoff condition.

Receivers find lies that are more consistent with the base-rate to be more plausible. We found that human receivers adjust which reports they call out based on the base-rate probability of sampling red marbles (Fig. 2.6). Collapsing over payoff conditions, in the 20% base-rate condition, the mean most accepted report was a report of 3.94 ± 0.22 red marbles. The mean accepted report was larger for higher base-rate conditions, with a report of 5.13 ± 0.17 in the 50% base-rate condition and 5.82 ± 0.24 in the 80% base-rate condition. The pairwise differences across all conditions were significant: at 20% vs. 50% ($z = 4.30$, $p < 0.0001$), 50% vs. 80% ($z = 2.40$, $p < 0.02$), and 20% vs. 80% ($z = 5.83$, $p < 0.0001$). This shows that receivers detect lies based on their consistency with statistical information they believe about the world.

Receivers are more likely to identify claims as lies when they are aligned with the reward structure. The human receivers' mode of accepted reports also differed depending on the payoff structure. When receivers were rewarded for blue marbles (rather than red) they tended to accept reports with more blue marbles, for all base-rate conditions. In the 20% base-rate condition, receivers' most accepted report was 2.52, 95% CI = [2.20, 2.84] red marbles when receivers were rewarded for blue marbles, and 5.36, CI = [5.07, 5.65] red marbles when receivers were rewarded for red marbles ($\bar{x}_d = 2.84$, $z = 12.93$, $p < 0.0001$). In the 50% base-rate condition, the receivers' most accepted report was 3.61, CI = [3.41, 3.81] red when receivers were rewarded for blue marbles, and 6.64, CI = [6.38, 6.90] red marbles when receivers were rewarded for red marbles ($\bar{x}_d = 3.04$, $z = 18.15$, $p < 0.0001$). Lastly, in the 80% base-rate condition, the receivers' most accepted report was 4.59, CI = [4.30, 4.88] red marbles when receivers were rewarded for blue marbles, and 7.05, CI = [6.69, 7.42] red marbles when receivers were rewarded for red marbles ($\bar{x}_d = 2.46$, $z = 10.41$, $p < 0.0001$; Fig. 2.6c). These results conclude that receivers call out lies by considering their alignment with the reward structure for both players.

2.2.3 Discussion

Experiment 1 tested several predictions of the `Recursive ToM` model in a dyadic lying game where players took turns reporting to an adversary the number of red marbles drawn from

a box, and classifying their adversary's reports as truths or lies. Critically, we manipulated the base-rate of red (vs. blue) marbles, as well as the payoffs associated with more red marbles for both players. We found support for three key predictions of the rational, theory of mind based model of lying and lie detection, namely that: (a) people lie more when the truth is less in their favor, (b) people choose lies based on their plausibility and payoffs, and (c) that people are also sensitive to plausibility and payoffs when detecting lies. In contrast, we found that lying behavior did not fit with the predictions of several alternative models.

While the `Equal Intrinsic Aversion` and `Unequal Intrinsic Aversion` heuristic models predict that people may ignore external payoff gains, and so should lie equally often regardless of the truth; people, in fact, do change how often they lie as a factor of the truth. Additionally, while both the heuristic models and the 0^{th} `Order ToM` model predict that the lies people do say will be insensitive to the base-rate; people, in fact, tune their lies based on what lies could be plausible. Lastly, while the `Null Hypothesis Significance Testing` model predicts that receivers, having no bias to prefer lies in a certain direction, should call out large and small reports equally often; people, in fact, consider the payoff structure when deciding how to call out lies. Under these considerations, the `Recursive ToM` models seems to accurately predict human lying and lie detecting behavior.

However, this experimental design cannot test a more subtle claim of the `Recursive ToM` model: that people tailor their lies to the beliefs they attribute to the receiver, even when those beliefs are different from their own. We test this claim in Experiment 2.

2.3 Experiment 2

Experiment 2 tested how people lie when senders and receivers have divergent beliefs about the probability of the world. We expand on the design of the lying game from Experiment 1: Now the distribution of marbles is only partially observable to the receiver, limiting their visual access. Importantly, the sender can observe what the receiver sees, but also has greater

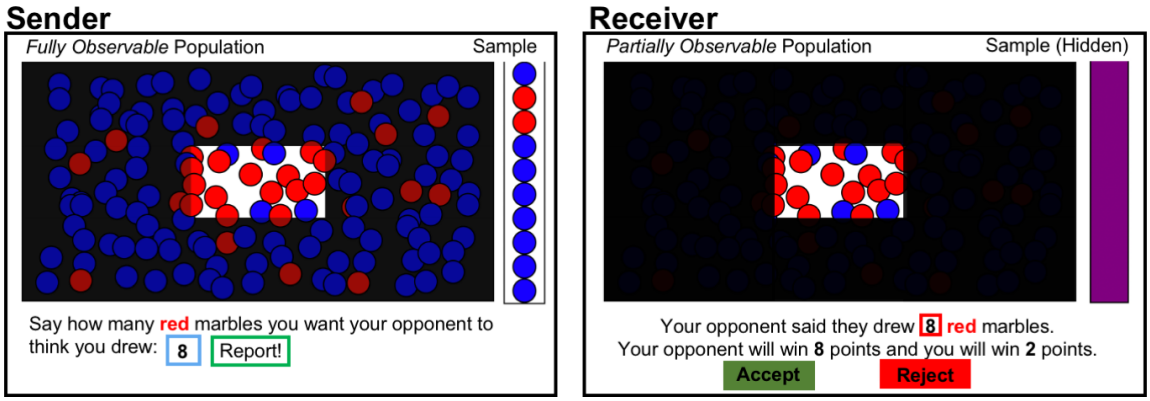


Figure 2.7. Experiment 2 used a partially observable dyadic lying game. For the sender, beliefs about the base-rate are fully observable: the sender knows the distribution of red and blue marbles observed by the receiver (in the inner white box), and the overall distribution (in the inner white and surrounding black box). For the receiver, beliefs about the base-rate are partially observable: the receiver can only observe the distribution of marbles in the window (the inner white box). Here, the sender believes the full population contains 20% red and 80% blue marbles, and they know the receiver observes a subset of the population that is 80% red and 20% blue marbles.

visual access, sometimes resulting in the sender having different beliefs about the base-rate than the receiver. This design serves to tease apart whether senders simply adjust their lies to *their own* beliefs or to beliefs about *their audience's* beliefs. Critically, a lying strategy that calls upon theory of mind to avoid detection ought to adapt lies to the audience's beliefs, and not simply to only one's own beliefs, in contrast with all accounts that generate lies without considering the listener. Thus, in Experiment 2, we address this question by dissociating the sender's and receiver's base-rate beliefs to investigate whether and how the receiver's beliefs influence a sender's lies.

2.3.1 Methods

Participants

291 participants were recruited from the undergraduate population at UCSD. Of these, 33 were excluded for failing to sufficiently answer at least 75% of the attention check questions. Therefore, 258 participants were included in our final data set.

Procedure

The lying game used in Experiment 2 resembled the game introduced in Experiment 1, except it separately manipulated the distribution of red and blue marbles in the box visible to the sender and the receiver (Fig. 2.7). In Experiment 1, the box of marbles was fully visible to both players; in Experiment 2, the box contained a window on one side (an inner white box) through which the receiver could see the distribution of red and blue marbles. The other side was open—the sender could see what the receiver saw through the window (the inner white box), as well as the full distribution of red and blue marbles (the inner white box and the surrounding black box). In other words, the population of marbles was fully observable for the sender, but only partially observable for the receiver. The sender could infer how the receiver's base-rate differed from their own, but the receiver had no information on which to evaluate whether the sender had a belief different from their own. As in Experiment 1, participants alternated between playing as the sender and receiver.

We used a 3×3 within-subject design manipulating: the sender's base-rate (total box; 20%, 50%, or 80% red); the receiver's base-rate (inner white box; 20%, 50%, or 80% red). This within-subject design necessarily required more participants, relative to Experiment 1. These conditions were randomly sampled for each trial.

To check whether our manipulation resulted in senders and receivers having divergent beliefs (as we intended), we asked participants to respond on a slider scale about the distribution of marbles from their own or their opponent's perspective (shown in Fig. 2.8). The left side of the slider bar was red and the right side was blue, so that the further rightward the bar was dragged, the more the bar was "filled in red." Labels below the slider ("more blue" to the left, "more red" to the right) helped to clarify the scale's direction. As in Experiment 1, participants also answered randomly distributed attention check questions about the number of red marbles drawn or reported. All participants received a total of 19 base-rate and attention check questions, except three subjects who received 18 (due to randomization).

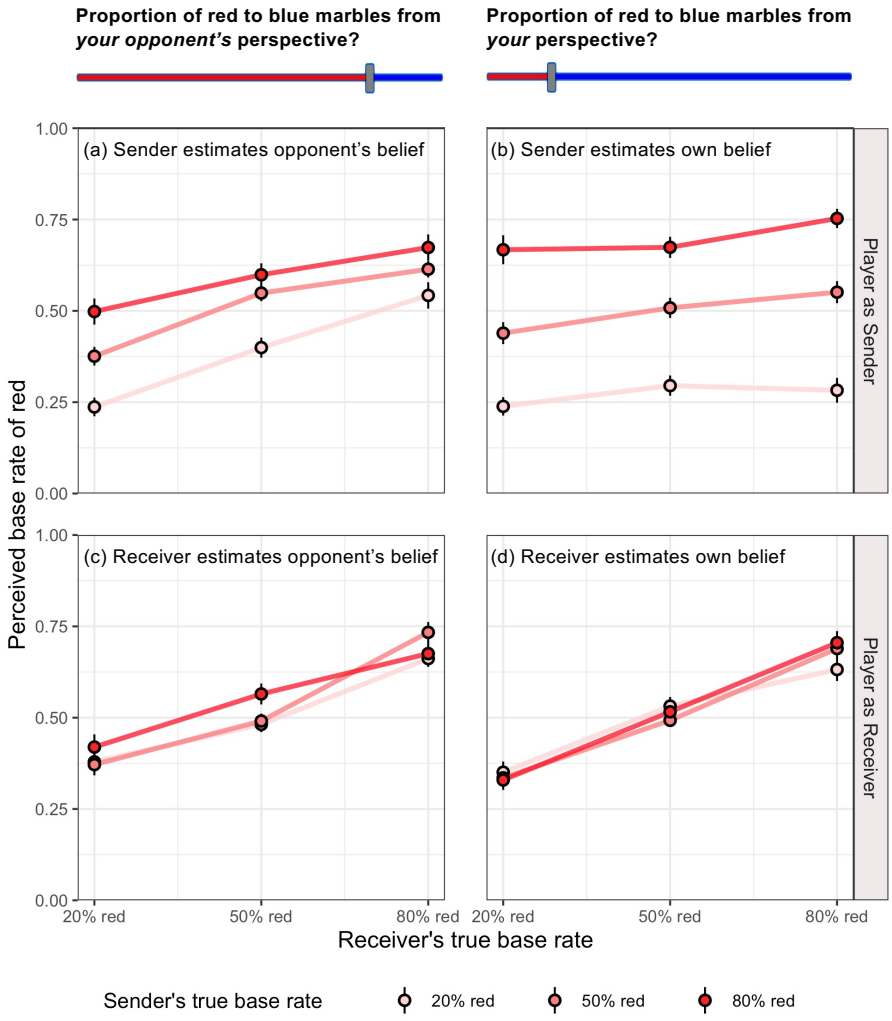


Figure 2.8. Experiment 2 participants' reported beliefs about the distribution of red and blue marbles. The x-axis is the receivers' base-rate condition, and the color is the senders' base-rate condition. The panel rows indicate the role of the participant as the sender (top) or receiver (bottom), and the panel columns indicate if the participant was asked about their opponent's (left) or their own beliefs (right). The y-axis, shows the participants' slider scale response. (a) When senders estimated their opponent's (receivers') beliefs, senders believed receivers' base-rate beliefs shifted with the receivers' true base-rate as expected, but surprisingly, the senders' true base-rate also had a small influence on their response. (b) When senders estimated their own (senders') beliefs, senders accurately assessed their own base-rate. (c) When receivers estimated their opponent's (senders') beliefs, and (d) when receivers estimated their own (receivers') beliefs, receivers responded the same.

In addition, unlike in Experiment 1 which manipulated the payoff structure of the game across conditions, Experiment 2 used only the red payoff condition's utility structure. In other

words, senders generally received points for more red marbles (and were motivated to over-report what they saw), and receivers received points for more blue marbles. Once again, participants played for a total of 100 trials.

2.3.2 Results

Manipulation Check

Did our manipulation of sender's beliefs, receiver's beliefs, and sender's beliefs about the receiver have the intended effects? For our manipulation to work, three conditions must be satisfied: (1) The sender must recognize that the receiver only has visual access to the distribution of marbles in the inner white box, and it guides the receiver's beliefs. (2) The sender must recognize that each player can hold different beliefs about the base-rate of marbles. (3) The receiver must actually believe the base-rate that they see, so as to make them susceptible to exploitation. We evaluated if participants' base-rate estimates (ranging from 0 to 100) varied as expected with player role (sender or receiver), question type (own or opponent's belief), and sender and receiver base-rate conditions.

Does the sender notice the receiver's base-rate? We checked if the study's key manipulation was successful—that the sender was aware of the receiver's beliefs, informed by the inner white box (Fig. 2.8a). A two-way ANOVA with an interaction revealed a significant effect of receiver base-rate ($F(6,636) = 17.06$, $p < 0.0001$), suggesting that sender understood that the receiver's beliefs about the base-rate were constrained by the aperture. There was also a significant effect of sender base-rate ($F(6,636) = 11.59$, $p < 0.0001$), indicating some "leakage" of the sender's beliefs into their assessment of the receiver's beliefs.

Does the sender believe the receiver has divergent beliefs? Our manipulations were specifically aimed to induce an asymmetry between the sender's beliefs about the receiver's beliefs (Fig. 2.8a) and the sender's own beliefs (Fig. 2.8b). We tested whether *whose* beliefs (sender or receiver) the sender was asked about interacted with the receiver and sender base-rate conditions, separately. For both interactions we found a significant effect (with receiver base-rate:

$F(2,1242) = 10.72$, $p < 0.0001$; with sender base-rate: $F(2,1242) = 21.74$, $p < 0.0001$). This means that our manipulations succeeded at separately influencing the sender's estimates of the base-rate, and their assessments of the receiver's beliefs about that base-rate.

Does the receiver assume the sender shares the same beliefs as themselves? Another assumption of the study is that receivers assume that the distribution of marbles visible to them approximately matches the distribution of marbles from which the sender is sampling. Alternatively, as the receiver, the participant may believe the game is rigged and distrust that the distribution for the players will match. To address this issues, we correlated the receiver's mean perceived base-rate beliefs about the sender's beliefs (Fig. 2.8c) and their own beliefs (Fig. 2.8d), and we found that the responses were positively correlated ($r = 0.74$, $t(7) = 2.94$, $p < 0.03$). This suggests that the receiver's belief about the sender's and their own beliefs approximately mapped onto each other for each condition, corroborating our assumption that the receiver defaults to assuming the sender's beliefs approximate their own.

In aggregate, our manipulations worked. The sender recognized that the receiver's beliefs about the base-rate were different from their own, and the receiver used their visible information to approximate the sender's likely beliefs.

Speakers design lies by considering receivers' prior beliefs

When do people lie? Under the `Recursive ToM` account, people should lie more when they believe the receiver has a higher base-rate belief, as having a high base-rate belief leaves people more susceptible to believing a large lie could be true (Fig. 2.9). In line with this, the sender's lying frequency increases with the true base-rate belief of the receiver ($\chi^2(2) = 138.5$, $p < 0.0001$), as well as the true base-rate experienced by the sender ($\chi^2(2) = 664.8$, $p < 0.0001$). These results imply that people can recognize when their audience is more exploitable, and they more frequently take advantage and lie in these situations.

How do people lie? Next, we examined how people chose to lie as a function of their own and the receiver's base-rate beliefs. As in Experiment 1, we extracted the centered intercept

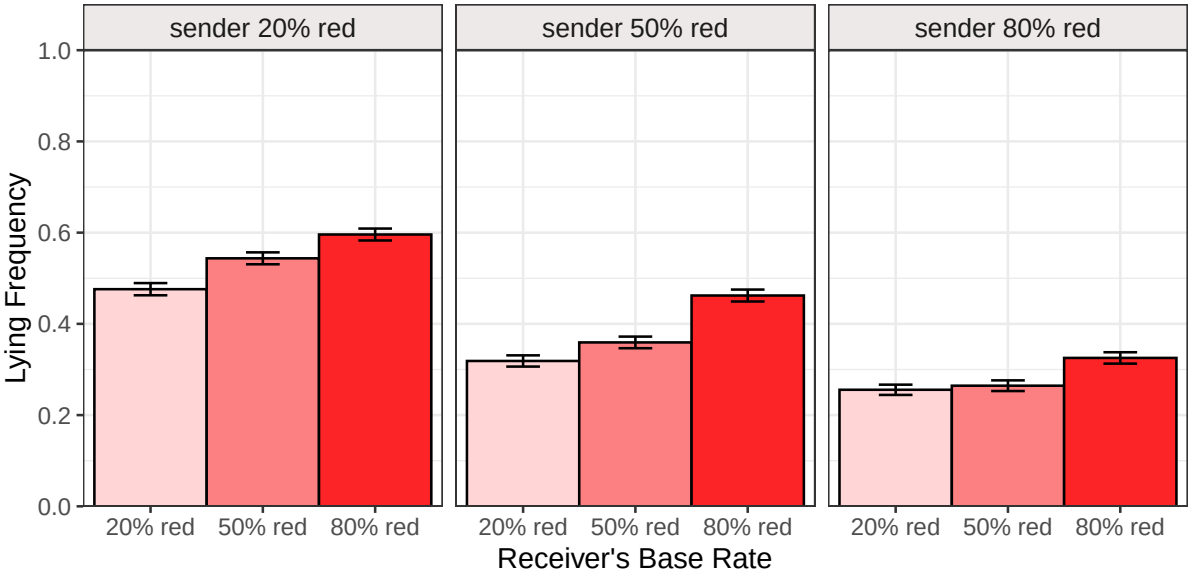


Figure 2.9. Experiment 2 rate of lying (as opposed to telling the truth) across conditions. There is as an effect of the receivers' (x) and the senders' base-rate condition (panels). People lie more when the receivers' base-rate belief is higher (e.g., 80%), suggesting that people recognize when their audience is more exploitable.

from the linear relationship between the true number of red marbles sampled and reported lies for each sender base-rate and receiver base-rate conditions. We then used this value as a summary of the sender's average lie in order to examine whether senders' or receivers' base-rate beliefs influenced people's lies, and to compare which was the stronger predictor (Fig. 2.10). The results revealed that both of the base-rate conditions were significant predictors of the reported lies, but the receivers' base-rate had a greater effect on lies ($\chi^2(12) = 1214.7$, $p < 0.0001$, $\hat{\omega}^2 = 0.119$) than the senders' base-rate ($\chi^2(12) = 34.7$, $p < 0.001$, $\hat{\omega}^2 = 0.003$). Thus, senders weighed receivers' prior beliefs *more* than their own when deciding how to lie. These results point to people's abilities to construct gain-increasing lies around the audience's unique beliefs and support the claim that senders are using an audience-based strategy to choose their lies.

2.3.3 Discussion

Experiment 2 sought to tease apart whether a sender designs lies that are solely influenced by the sender's own beliefs, or their ToM-driven reasoning about the receiver's beliefs. The latter

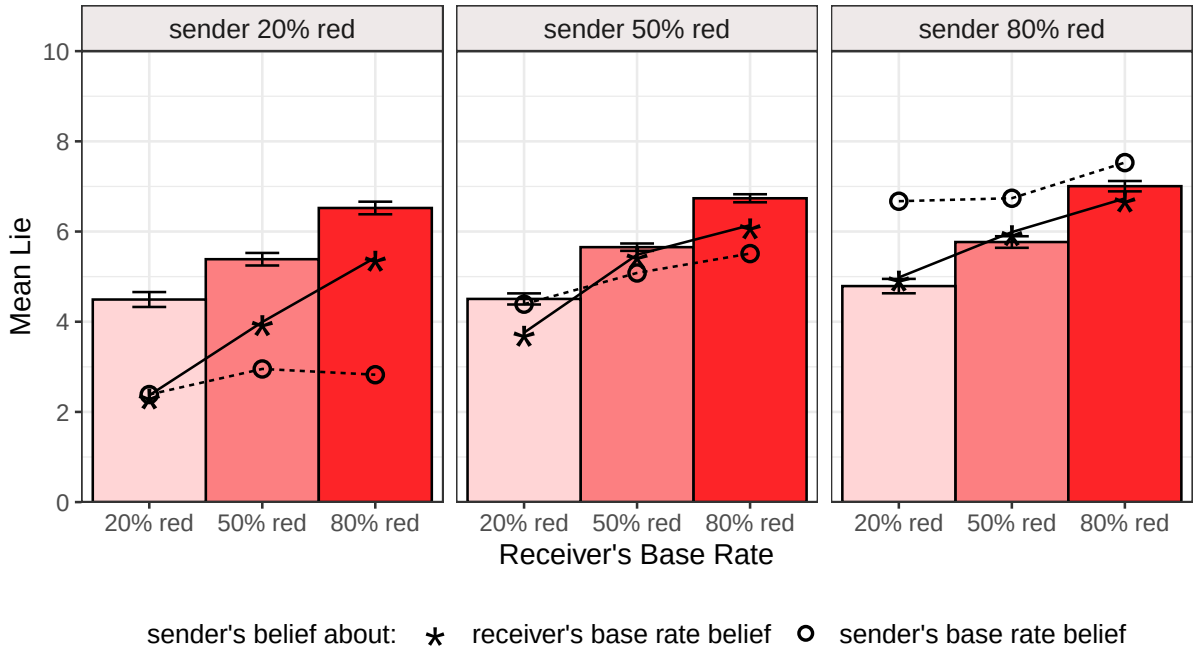


Figure 2.10. Experiment 2 average lie across conditions, computed from the intercept of the linear fit. There is a strong effect of the receivers' base-rate condition (x-axis), and little effect of the senders' base-rate condition (panels). Stars represent the senders' estimates of the receivers' belief about the base-rate (from Fig. 2.8a), and circles represent the sender's direct estimate of the base-rate (from Fig. 2.8b). The average lie more closely tracks senders' estimates of receivers' beliefs, suggesting that senders use theory of mind to choose how to lie.

is qualitatively predicted by the `Recursive ToM` account of lying, which uniquely considers the beliefs of the receiver. Thus, the partially observable lying game allowed us to evaluate the role of ToM, by considering how people lie when there is an explicit mismatch between their own prior beliefs and their estimates of the receiver's prior beliefs. In these settings, we found that peoples' lies are better predicted by beliefs about the receiver, as opposed to beliefs about themselves, further supporting a critical role of theory of mind in deciding how to lie.

2.4 General Discussion

The current work presents a unified framework underlying lie design and detection, formalized as recursive social reasoning. This approach highlights how liars and lie detectors plan their behaviors via interactive, adversarial reasoning: Senders design lies by infer-

ring the likelihood the receiver detects potential lies; receivers detect lies by inferring if and how the sender would lie. We compared our `Recursive Theory of Mind (ToM)` account to three accounts that do not require ToM, namely the `Equal Intrinsic Aversion Heuristic`, `Unequal Intrinsic Aversion Heuristic`, and 0^{th} `Order ToM` accounts. Compared to the other models, `Recursive ToM` uniquely generated several key diagnostic qualitative predictions about patterns in lying and lie detecting behavior: (1) people should lie more when the truth is less favorable, (2) people should balance payoff gains and plausibility when deciding what lie to say, (3) people should cater their lies to what they think their audience will find plausible, and (4) when detecting lies, people should be sensitive to plausibility and payoffs, as well.

We empirically tested and showed that our model explained the rate and content of lies people produced, and which lies they detected. In Experiment 1, people lied more when the truth was less favorable, consistent with lying being strategically scaled to circumstances, rather than being a small constant offset if lying was based simply on anchoring to speaker's knowledge of the truth. Furthermore, people produced larger lies when those larger lies were more consistent with the base-rate, indicating that they balanced payoff and plausibility, in contrast with the idea that lies are tempered largely by moral considerations. Finally, people detected lies by being sensitive to payoff and plausibility, indicating that lie detectors are sensitive to the content of the lie, rather than solely considering superficial cues. Experiment 2 provides stronger diagnostic evidence for a role of theory of mind in lying: when senders and receivers have a mismatch in beliefs about the world, senders tuned their lies to the audience's beliefs more than to their own beliefs. This further confirms that lies are crafted to balance payoff and plausibility for the listener. Altogether, we take these results as evidence that people can and do spontaneously calibrate their lying and lie detecting by employing theory of mind.

The idea that people reason about others' minds when strategizing about lies builds on evidence from previous work. For instance, children's theory of mind development is linked to their ability to lie and to maintain their lies over time (Ding et al., 2015; K. Lee, 2013; Talwar et al., 2007). In adults, beliefs about the receiver's level of suspicion predicts whether or not

people choose to deceive (Franke et al., 2020; Gneezy et al., 2018; Montague et al., 2011; Nagin & Pogarsky, 2003; Ransom et al., 2019; Rogers et al., 2017). However, the current results support a stronger claim: theory of mind underlies a unified model of lying and lie detection—the frequency *and* magnitude of lies are calibrated using people's interactive social reasoning. People rationally lie and detect lies, using social reasoning to predict other agents' behavior.

The view of lying and lie detection supported by our experiments—as strategic acts driven by theory of mind reasoning—adds nuance to the prevailing focus in the literature. Previous research on lying often considered high-stakes, hard to detect, lies—such as in tax evasion and fraud; or during police interrogations—and asked distinct questions, such as whether highlighting the moral salience of lying could increase honesty (Kristal et al., 2020; Mazar et al., 2008), or whether superficial behaviors could be used to detect lies (DePaulo et al., 2003). This work emphasized that lies can be constrained by intrinsic morals (Mazar et al., 2008), and that people are poor at performing lie detection in some circumstances (Bond & DePaulo, 2006), in part because they are too automatically trusting to succeed (Levine, 2014). In contrast, here we focus on common, everyday lies—lies that are commonplace but where extreme lies are easily detectable, analogous to overstating your resume qualifications, or lying about your height on dating profiles. We show that theory of mind shapes the generation and detection of these common, everyday lies. While we agree that moral considerations (for example) can modulate lies, and that there are limits to people's lie detection abilities, the current work suggests that theory of mind plays a foundational role in lie generation and detection, acting in concert with these additional factors. We expect this is true of high-stakes lies as well.

In real-world settings, this strategic theory of mind strategy and a moral individually-focused lying aversion are not mutually exclusive. People may lie, for example, by primarily trading off maximizing their gain and avoiding audiences' detection, but they may secondarily avert conventionally unethical lies. Both cognitive mechanisms are likely weighed variably across contexts, which may partly explain why the propensity to lie varies across experimental paradigms and laboratory versus field studies (Gerlach et al., 2019). Our results indicate that

human lying behavior is not driven solely by individual factors, but instead takes into account what others are likely to think about potential lies. However, this does not mean that there is no role for individual factors—at the very least, there is likely to be individual variation in aversion to lying, even though lies, when told, are strategically designed for the audience. Future work may more directly compare how other factors, like moral reasoning, trades off with audience-related factors.

Our results relate to the literature on recursive reasoning in behavioral game theory. Classic work in this domain classifies agents as level-k reasoners (Crawford & Iriberri, 2007; Stahl, 1993), or describes their reasoning capacity in terms of a cognitive hierarchy of recursion depths (Camerer et al., 2004). Work in this field has attempted to characterize peoples' exact recursion distribution, using games designed explicitly to measure the number of levels of recursion each person has computed—such as the p-beauty contest (Ho et al., 1998; Nagel, 1995). This work shows that people are well-characterized by an average recursion depth of 1.5 (Camerer et al., 2004). How many levels of recursion do people compute in adversarial communication contexts, when lying or detecting lies? Our experiments demonstrate that in adversarial communication contexts, listeners and speakers are both at least level-2 reasoners: Listeners consider the goals of speakers when detecting lies, and speakers consider the beliefs of listeners when designing lies. This provides a lower bound on recursion depth of participants in our experiment. However, our data cannot establish an upper bound, or participants' precise k-level. Future work should adapt finer-grained methods to identify the level of recursion that people employ during adversarial communication.

We intentionally set up our experiments to resemble the common situation in which larger magnitude lies are both more rewarding if accepted, and also easier to detect, to create simple countervailing pressures on liars. This situation occurs for many real-world situations in which numbers are reported—fraud about balance sheets, taxable income, and revenues, for example. All have the property that lie magnitude monotonically increases reward (if believed), but also monotonically increases the risk of detection. However, detectability may not always

trade off with value to the sender: For example, imagine the subtle yet highly advantageous lie possible when a test is graded pass/fail, and a students' score is only one percentage point from the threshold. In this case, fudging the score by only one point results in a change from no credit, to full credit. More generally, this situation arises when the utility (value) of an answer does not scale linearly with the possible responses. We expect that the same kind of recursive theory-of-mind based reasoning seen here would be used in this more complex case. In other words, when both the sender and receiver are aware of this non-linear distribution of utility, their lie detection should adjust to consider responses more likely to be lies when they are more in line with the senders' goals, even though this consideration is more complex to consider than simply larger numbers equating to higher utility. Future work may test whether people employ recursive social reasoning when faced with more complex, non-linear distributions of utility.

The current experimental setting also concerns a situation in which accurately detecting and calling out a lie is advantageous. While this is true in many real-world contexts, lying interactions are often more complex. For example, lie detection may be strategically concealed. There are many settings where, upon detecting a lie, the most prudent action is to not reveal that the lie has been detected (perhaps due to some utility cost to revealing private information, or for instigating conflict). This strategic concealment adds another deception to the situation. In this situation, the receiver faces a decision: Whether to tell the truth and state that a lie was detected; or whether to lie by omission, and choose not to reveal that a lie was detected. More broadly, agents' decisions to show or hide knowledge about other agents' goals and beliefs likely plays a crucial role in the arms race between deceivers and detectors. Characterizing the reasoning underlying strategic concealment of lie detection is an important next step in expanding the scope of our model, toward understanding the natural complexity of adversarial communication.

Like all laboratory experiments, ours were designed to isolate and measure particular effects: In the research we present here, we show that people can rationally lie by considering their opponent's beliefs and goals. Our experiments used a low-stakes, game-like setting, potentially increasing subjects' willingness to lie, and to lie strategically, as compared to high-stakes

real-world communication. Similarly, lying in our experiments was compartmentalized from subjects' real lives, thus eliminating considerations of reputation management and downstream reciprocity. By having participants play against an artificial agent, we can control how opponents behave, to more effectively measure how people respond; but surely incentives will differ in this setting compared to face-to-face communication. Furthermore, our results pertain to the behavior of many individuals, compared in between-subject conditions in Experiment 1, or aggregated in Experiment 2—in some cases group behavior may appear to fit a particular model, while individuals do not (Goodman et al., 2008; Vul et al., 2014). Future work should examine individual differences in reasoning, to evaluate the extent to which key recursive reasoning patterns apply to each participant considered in isolation.

Lastly, by alternating roles on consecutive game rounds, participants may have become more likely to consider how their counter-party would respond to their behavior, relative to situations in which they had never experienced being the counter-party. This would be in line with developmental evidence for the role of first-person experience in some aspects of action understanding (e.g., Gerson & Woodward, 2014). Furthermore, repeatedly taking turns between roles seems to mirror the level-n recursive theory of mind reasoning described in our model. How might the act of alternating roles facilitate recursive reasoning? Future work should explore to what extent first-person experience as both the sender and receiver facilitates or is necessary for reasoning about others' beliefs, goals, and actions in the context of lying and lie detection. In the context of our task, this could involve having participants play only one role, or presenting roles in blocks of trials rather than in alternation.

Overall, we provide evidence that people can spontaneously calibrate their lying and lie detecting by employing theory of mind. Incorporating mental state reasoning into the understanding of lies may be a useful path toward smarter, more human-like AI to automate the detection of false information ("fake news") on social media. False claims are highly prevalent online, and motivated by particular goals; but current AI systems do not incorporate others' likely knowledge and motives, limiting lie detection. The current data suggest that a greater emphasis

on socially intelligent mechanisms is warranted in our push towards more epistemically vigilant AI systems (Sperber et al., 2010). Overall, our work both advances basic science of cognition and mental state reasoning, and moves toward an automated system for improved detection of false claims.

2.5 Acknowledgments

This material is based upon work supported by the NSF Graduate Research Fellowship under Grant DGE-1650112 to Lauren A. Oey and the National Science Foundation Grant BCS-1749551 to Adena Schachner. Previous versions of this work have been presented at the Cognitive Science Society and Society for Philosophy and Psychology conferences (Oey et al., 2019a, 2019b; Oey & Vul, 2021).

Chapter 2 is a reprint of material as it appears in Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, *152*(2), 346-362. The dissertation author was the primary investigator and author of this paper.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115–1153.

Allcot, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, *17*(10), 841–844.

Becker, G. S. (1968). Crime and punishment: An economic approach. In *The Economic Dimensions of Crime* (pp. 13–68). Palgrave Macmillan.

Blair, J. P., Levine, T. R., & Shaw, A. S. (2010). Content in context improves deception detection accuracy. *Human Communication Research*, *36*(3), 423–442.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234.

Bond, C. F., Howard, A. R., Hutchinson, J. L., & Masip, J. (2013). Overlooking the obvious: Incentives to lie. *Basic and Applied Social Psychology*, *35*(2), 212–221.

Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, *71*, 499–515.

Bruer, K. C., Zanette, S., Ding, X. P., Lyon, T. D., & Lee, K. (2019). Identifying liars through automatic decoding of children's facial expressions. *Child Development*.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory*, *6*(3), 203–242.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, *119*(3), 861–898.

Capraro, V., Schulz, J., & Rand, D. G. (2019). Time pressure and honesty in a deception game. *Journal of Behavioral and Experimental Economics*, *79*, 93–99.

Crawford, V. P., & Iriberri, N. (2007). Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica*, *75*(6), 1721–1770.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*(2), 238–257.

Dennett, D. (2009). Intentional systems theory. In *The Oxford Handbook of Philosophy of Mind* (pp. 339–350).

DePaulo, B. M., Malone, B. E., Lindsay, J. J., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118.

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-mind training causes honest young children to lie. *Psychological Science*, *26*(11), 1812–1821.

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, *32*(1), 88–106.

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, *54*(3), 414–420.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, *58*(4), 723–733.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), e5738.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise–an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547.

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

Franke, M., Dulcinati, G., & Pouscoulous, N. (2020). Strategies of deception: Under-informativity, uninformativity, and lies — misleading with different kinds of implicature. *Topics in Cognitive Science*, *12*(2), 583–607.

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*(1), 1–44.

Gerson, S. A., & Woodward, A. L. (2014). Learning from their own actions: The unique effect of producing actions on infants' action understanding. *Child Development*, *85*(1), 264–277.

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological Science*, *20*(3), 393–398.

Gladwell, M. (2019). *Talking to Strangers: What We Should Know About the People We Don't Know*. Little, Brown; Company.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*(1), 384–394.

Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, *108*(2), 419–453.

Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior and Organization*, *93*, 293–300.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Granhag, P. A., & Strömwall, L. A. (2002). Repeated interrogations: Verbal and non-verbal cues to deception. *Applied Cognitive Psychology*, *16*(3), 243–257.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics Vol. 3: Speech Acts* (pp. 64–75). Academic Press.

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press.

Hancock, J. T., & Toma, C. L. (2009). Putting your best face forward: The accuracy of online dating photographs. *Journal of Communication*, *59*(2), 367–386.

Hilbig, B. E., & Hessler, C. M. (2013). What lies beneath: How distance between truth and lies drives dishonesty. *Journal of Experimental Social Psychology*, *49*, 263–266.

Ho, T.-H., Camerer, C., & Weigelt, K. (1998). Iterated dominance and iterated best response in experimental "p-beauty contests". *The American Economic Review*, *88*(4), 947–969.

Hurkens, S., & Kartik, N. (2009). Would I lie to you? On social preferences and lying aversion. *Experimental Economics*, *12*(2), 180–192.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

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

Kraut, R. E. (1978). Verbal and nonverbal cues in the perception of lying. *Journal of Personality and Social Psychology*, *36*(4), 380–391.

Kristal, A. A., Whillans, A. V., Bazerman, M. H., Gino, F., Shu, L. L., Mazar, N., & Ariely, D. (2020). Signing at the beginning versus the end does not decrease dishonesty. *Proceedings of the National Academy of Sciences*, *117*(13), 7103–7107.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, *117*(3), 785–807.

Lee, K. (2013). Little liars: Development of verbal deception in children. *Child Development Perspectives*, *7*(2), 91–96.

Levine, T. R. (2010). A few transparent liars: Explaining 54% accuracy in deception detection experiments. *Annals of the International Communication Association*, *34*(1), 41–61.

Levine, T. R. (2014). Truth-Default-Theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*(4), 1–15.

Levine, T. R. (2019). *Duped: Truth-Default Theory and the Social Science of Lying and Deception*. University of Alabama Press.

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

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley.

Mann, S., Vrij, A., & Bull, R. (2004). Detecting true lies: Police officers' ability to detect suspects' lies. *Journal of Applied Psychology*, *89*(1), 137–149.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

Montague, R., Navarro, D. J., Perfors, A., Warner, R., & Shafto, P. (2011). To catch a liar: The effects of truthful and deceptive testimony on inferential learning. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 1312–1317).

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, *85*(5), 1313–1326.

Nagin, D. S., & Pogarsky, G. (2003). An experimental investigation of deterrence: Cheating, self-serving bias, and impulsivity. *Criminology*, *41*(1), 167–194.

Oey, L. A., Schachner, A., & Vul, E. (2019a). Designing good deception: Recursive theory of mind in lying and lie detection. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 897–903).

Oey, L. A., Schachner, A., & Vul, E. (2019b). Recursive theory-of-mind in the design of deception: A rational model of lying and lie detection.

Oey, L. A., & Vul, E. (2021). Lies are crafted to the audience. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (pp. 791–797).

Ohtsubo, Y., Masuda, F., Watanabe, E., & Masuchi, A. (2010). Dishonesty invites costly third-party punishment. *Evolution and Human Behavior*, *31*(4), 259–264.

Phillips, J., Ong, D. C., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás and Endress (2010). *Psychological Science*, *26*(9), 1353–1367.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, *4*, 515–526.

Ransom, K., Voorspoels, W., Navarro, D. J., & Perfors, A. (2019). Where the truth lies: How sampling implications drive deception without lying. *PsyArXiv*.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*(3), 338–342.

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead other. *Journal of Personality and Social Psychology*, *112*(3), 456–473.

Schelling, T. C. (1960). *The Strategy of Conflict*. Harvard University.

Serota, K. B., Levine, T. R., & Boster, F. J. (2010). The prevalence of lying in America: Three studies of self-reported lies. *Human Communication Research*, *36*(1), 2–25.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justification). *Psychological Science*, *23*(10), 1264–1270.

Shalvi, S., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011). Ethical manoeuvring: Why people avoid both major and minor lies. *British Journal of Management*, *22*, 16–27.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Stahl, D. O. (1993). Evolution of smart$_n$ players. *Games and Economic Behavior*, *5*(4), 604–617.

Street, C. N. (2015). ALIED: Humans as adaptive lie detectors. *Journal of Applied Research in Memory and Cognition*, *4*(4), 335–343.

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428–453.

Talwar, V., Gordon, H. M., & Lee, K. (2007). Lying in the elementary school years: Verbal deception and its relation to second-order belief understanding. *Developmental Psychology*, *43*(3), 804–810.

Toma, C. L., Hancock, J. T., & Ellison, N. B. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin*, *34*(8), 1023–1036.

Tyler, J. M., Feldman, R. S., & Reichert, A. (2006). The price of deceptive behavior: Disliking and lying to people who lie to us. *Journal of Experimental Social Psychology*, *42*, 69–77.

Vrij, A. (2008). Nonverbal dominance versus verbal accuracy in lie detection: A plea to change police practice. *Criminal Justice and Behavior*, *35*(10), 1323–1336.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, *10*(4), 141–142.

Vrij, A., Hartwig, M., & Granhag, P. A. (2019). Reading lies: Nonverbal communication and deception. *Annual Review of Psychology*, *70*, 295–317.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Withall, A., & Sagi, E. (2021). The impact of readability on trust in information. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (pp. 2370–2376).

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology*, *14*, 1–59.

Chapter 3

Accurate approximations about the truth from literally false messages

Communication can be weaponized to manipulate others' beliefs, most glaringly via explicit lies. We investigate one defense mechanism: people can infer the truth from false messages when they expect that (1) speakers have adversarial motives that direct their lies and (2) bigger lies are costlier. We show in a lab experiment that people can correct for bias in lies when these conditions are satisfied, but with decreased precision. When people adjust what information they glean from expected dishonesty, how might this perturb dyadic, and moreover collective, communication channels? Through probabilistic simulations, we find that deceptive communication systems converge to equilibrium states, in which listeners extract accurate (but less precise) estimates of the truth. Furthermore, when listeners correct for messages assuming that they are distorted, even cooperative speakers (who want listeners to have the correct interpretations) should lie. Liars do not get their way, but they make communication noisier for listeners and other speakers.

deception; communication; lie detection; social cognition; probabilistic models

Communication generally aims to faithfully transmit information from a speaker's mind to a listener's mind. Typically, the listener expects the speaker to be honest, and the speaker expects to be interpreted as honest (Grice, 1975). The aligned goals allow for people to effortlessly converge on a shared communication system (Clark, 1996; Tomasello et al., 2005). However, there are many forms of communication that feature misaligned goals and call upon listeners to be vigilant, such as in deception (Sperber et al., 2010). If cooperative communication aims for the listener to infer an accurate depiction of what the speaker thinks, one aim of deceptive communication is to induce in the listener a distorted depiction. Broadly, this highlights information transmission as a key incentive for both cooperative and deceptive communication.

Consider a speaker who strategically lies – in doing so, they want to distort the listener's belief, rather than simply trying to remain undetected. Similarly, the listener wants to extract an accurate representation of the truth from the distorted message, rather than simply sleuthing out whether a message is a lie or not. Thinking of deceptive communication as information transmission couches listeners as lie *interpreters*, asking people what meaning they extract out of a message. This perspective deviates from the traditional focus of listeners as lie *detectors*, who categorize messages as true or false e.g., Bond and DePaulo, 2006, 2008; Leach et al., 2004; Levine et al., 1999; Oey et al., 2023; ten Brinke et al., 2016. Here, we examine inference when deception is already suspected, so listeners are not burdened with worrying about whether a message is a lie – rather, what is the truth, given that this message is likely a lie. In doing so, we introduce a formulation of the goals of deception that emphasize a *social intention* to manipulate others' beliefs.

The premise of transmitting distorted messages occurs across numerous human communication systems. One such communication system is letters of recommendation, in which letter writers seek to promote their candidate, so they are highly motivated to inflate their candidate's apparent qualifications. However, they face constraints — for example, baldfaced over-embellishments may hurt letter writers' reputation. Meanwhile, letter readers want to accurately assess the candidate's qualifications. The asymmetric goals between letter writers and

readers promotes the passing of distorted messages. Qualified candidates are not simply "good," they are "the complete package."

Given that letters of recommendation are fraught with embellished language, the communication system at first glance seems prone to erroneous information transmission, like how learners who are exposed to biased samples tend to draw biased inferences (Feiler et al., 2013; Hogarth et al., 2015). Yet, our continued use of letters of recommendation superficially suggests that letters are in large scale effective at communicating information. And for individual readers to value using letters, they must expect to extract meaningful information. Just as learners who are aware of the sampling constraints correct their generalizations (Hayes et al., 2019), perhaps readers systematically correct for biased language in letters. This raises a puzzle at the level of dyadic communication: how do people interpret distorted messages?

On one hand, people may rely on domain-specific, established conventions that give rise to the meaning of messages (Lewis, 1969). A distorted message serves as one arbitrary solution to the coordination problem on meaning. Both speakers and listeners align on mapping the message ("the complete package") onto the interpretation ("good"). However, this serves as a dissatisfying explanation to how distorted communication systems arise because a seemingly more salient solution would be to transmit truthful messages (Lewis, 1969; Schelling, 1960), as opposed to distorted ones.

Alternatively, people may interpret distorted messages guided by domain-general mechanisms, namely rational theory-of-mind reasoning (Oey et al., 2023). The core mechanism driving people's lie interpretation is an assumption about speakers' goals and an intuitive understanding of how goals drive speakers' behavior. Broadly, rational theory-of-mind frameworks are grounded on the assumption that both speakers and listeners generate decisions that maximize their rewards, and they intuit that other agents do the same (Baker et al., 2017; Jara-Ettinger et al., 2016). Recent implementations have proven useful for explaining numerous speaker-phenomena in deception, such as how suspicion influences people to preferentially mislead or be uninformative (Franke et al., 2020; Ransom et al., 2019), or how plausibility drives the extremeness of lies

(Oey et al., 2023). In interpreting communicative messages too, listeners draw meaning about what was said guided by their assumptions of speakers' goals. When listeners' assumptions about speakers are misplaced, listeners are vulnerable to being deceived. For example, in pedagogy, learners assume that knowledgeable teachers demonstrate a concept by being fully informative. If teachers omit information, learners can be misled into thinking they know all that there is to know about the concept, when this is in fact a false conclusion (Bonawitz et al., 2011). When made aware of the teacher's tendency to omit information, even children can adjust what meaning they draw from messages, such as "there is more to learn about this concept than what I have been told" (Gweon et al., 2014). Therefore, listeners seem to be equipped with cognitive tools to be vigilant to what information they receive.

My work proposes that vigilant listeners can make richer inferences than those seen in the previous literature. Ushered by rational assumptions, a listener, who observes a speaker's distorted message, may reverse engineer *what the speaker thought was the truth*. A key prediction of this framework is that people can robustly tune how they interpret the truth to their knowledge about the speaker's goals. Back to the teacher example, not only might listeners infer that there is more to learn, rather more specifically, sufficient knowledge about the speaker will invite listeners to infer that there is *one* more thing to learn, such as the teacher withholding information to test generalization about a concept. In contrast, learners could also suspect that there is *many* more things to learn because the teacher is withholding a lot of information, as to encourage learners to explore more independently. Critically, rational theory-of-mind does not automatically bestow omniscience – the accuracy of listeners' inferences about the truth depends on the veracity of their mental model of the speaker, such as how they conceptualize the speakers' rewards and costs to lie.

For distorted communication systems to arise, we presume that there are two necessary preconditions that constrain how goals influence messages. (1) Speakers want to induce the listener into believing something about the world that is not only literally false, but is also favorable to the speaker. For listeners to tune their inferences, they need to be aware of the

speaker's broad goals. (2) Speakers must face costs to constructing lies that deviate from the truth. For instance, process models, such as those that propose a direct relationship between speakers initially thinking about the truth and secondarily manipulating the truth to produce a lie e.g., Debey et al., 2014; Walczyk et al., 2014, broadly predict that lies that more distantly deviate from reality require more cognitive effort to construct. These costs might be due to increasing risk of detection (Oey et al., 2023), loss of plausible deniability (Pinker et al., 2008), higher cognitive load (van't Veer et al., 2014), managing reputation (Abeler et al., 2019), moral beliefs (Mazar et al., 2008), or social norms (Gino et al., 2009).

Deploying a novel behavioral task, we tested if people successfully infer the truth when the two preconditions are met. A rational theory-of-mind framework not only predicts that people should succeed with sufficient knowledge about the speaker, but it uniquely predicts that people tune their truth inferences to different beliefs about speakers' goals and costs. Participants played a game in which a sender draws red and blue marbles from a jar and sends a manipulated representation of their marbles to a judge by clicking marbles on the interface (a physical cost). Seeing the manipulated representation, the judge guesses how many red marbles were truly drawn. We found that people robustly interpreted distorted messages in a way that was tuned to speakers' directional goal and the magnitudinal cost.

Applying probabilistic models, we tested the downstream consequences of listeners' lie interpretations. The rational theory-of-mind framework sets up a first principles approach to simulate how agents' goals influence distorted communication systems. One possibility is that speakers (and listeners) plan their behaviors attempting to out-strategize the other. The result is an arms race between speakers ratcheting up the extremeness of their distorted messages and listeners ratcheting up their corrections, so that "letters of recommendation" become increasingly detached from reality. Contrary to this intuition, by applying probabilistic simulations, we showed that messages and interpretations converge to an equilibrium state when listeners suspect speakers' goals. Rewards and costs of speakers influence the accuracy and precision of listeners' truth inferences: as the cost to lying decreases relative to the reward for deceiving a listener,

the transmission of the truth remains unbiased (though more imprecise). Furthermore, when listeners generalize their suspicion to others, the consequence is that other speakers are indirectly affected. Cooperative speakers, wanting to guide listeners to accurate beliefs, are pressured to say *dishonest* messages when they expect vigilant listeners.

Overall, our study informs our scientific understanding of how distorted messages, but nonetheless faithful transmission, can perpetuate in communication systems. Our probabilistic, goal-based paradigm reveals that distorted communication systems are not simply odd pockets of anomalies, rather they commonly occur and produce systematic behaviors. Underlying these communication systems are people's robust ability to engage in rational theory-of-mind reasoning, which powers people to extract clairvoyant insight about the truth from falsehoods.

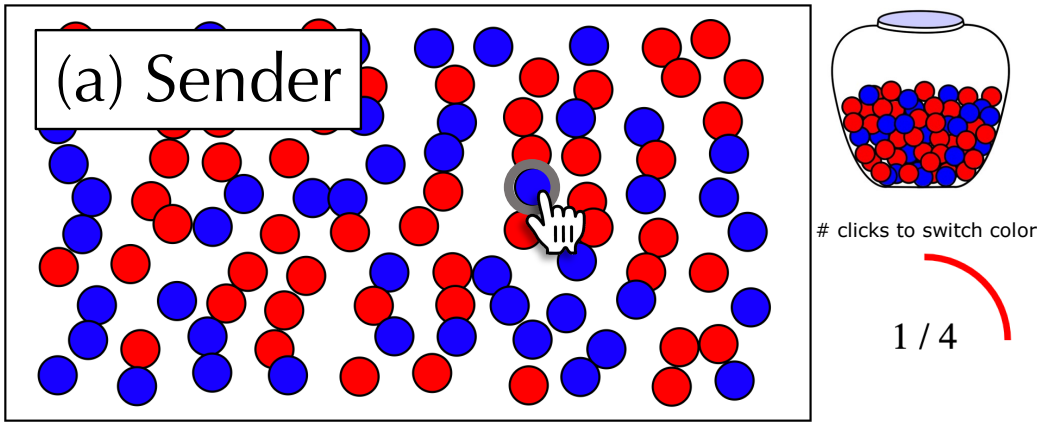
3.1 Human Experiment: Testing Goals and Costs as Preconditions to Infer the Truth

As an initial proof of concept, we tested if people not only infer the truth from lies, but they robustly tune their inferences to the speaker's goals and costs, in an experimental setting where the preconditions apply. Namely, that they are aware of the speakers' motive to directionally bias their lie, and that they face costs for producing more extreme lies.

3.1.1 Methods

Participants

Participants were recruited from the undergraduate population at University of California, San Diego to participate in an online game for course credit. Data was collected from 254 participants. Of these, 44 participants were excluded for failing to answer at least 75% of the attention check questions within a ± 5 error, and six participants produced multiple responses that were out-of-bounds. Participants who produced a single out-of-bounds trial had that trial excluded from analysis, but their remaining trials were included. In total, 204 subjects were included in our final data set. Informed consent was obtained from all participants, and the study



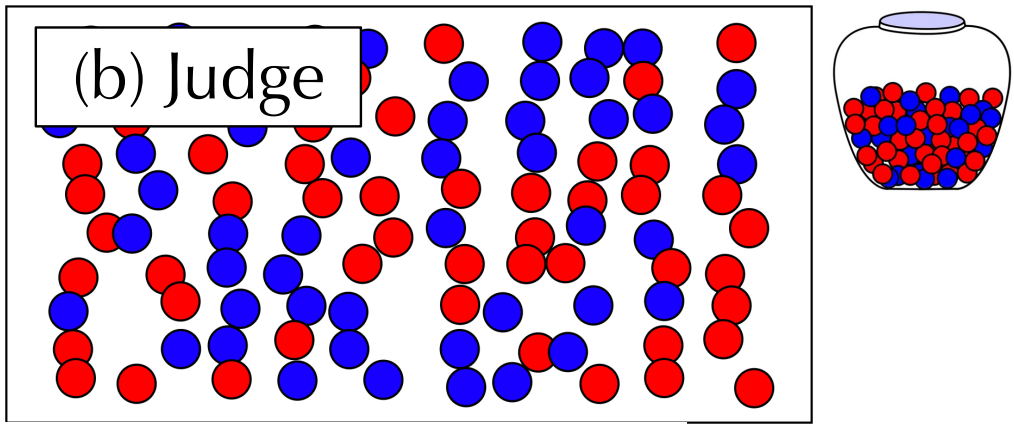


Figure 3.1. Inferring truth game design. (a) The sender sampled marbles, and could manipulate what they showed their opponent about how many red marbles they drew by clicking marbles in the display to flip their color. The sender is told in text how many of each color marble they originally drew (e.g., "You drew 48 red...") and how many they would currently report based on their clicks (e.g., "You will tell your opponent you got 51 red..."), and a progress bar shows how many more clicks are needed to switch the next marble. (b) The judge tried to estimate how many marbles the sender truly drew from what the sender reported. In this example, the sender wants the judge to *Overestimate*, and producing larger lies follows a `Quadratic` cost function (requires additional click for each additional flip). Here, the *truth* was 48 red marbles, but the *message* was a lie of 51. The judge *estimated* the truth to be 50.

was approved by the university's Institutional Review Board.

Procedure

Participants played an adversarial communication game, alternating between the roles of sender and judge. Senders saw a display of 100 red and blue "marbles" arranged in a 2D jittered grid, reflecting the ground *truth* sampling of red and blue marbles from a virtual jar (Figure 3.1). The sender could alter how many red (and blue) marbles were in the display by manually clicking individual marbles to swap their color, before sending to the judge the *message*, the altered snapshot of the marbles. The judge, in turn, sees the shaken display of marbles and the number of red marbles in it (i.e., the message), and then has to *estimate* the original, ground-truth number of red marbles.

The players' goal was to win against the other player by the largest possible point differential. Judges lost points corresponding to the absolute (L1) error of their estimate, so in Figure 3.1, a guess of 50 when the truth was 48 resulted in -2 points. Meanwhile, senders gained points for the judge's error in the direction of the sender's goal, so a sender who wanted the judge to *overestimate* got $+2$ points. If the judge guessed in the opposite direction (e.g., underestimated instead), the sender got 0 points, but the judge still got -2 points for their absolute error.

The critical between-subject manipulations were the *goals* and *costs* of deception for the sender. Senders were assigned the goal to make the judge either *Overestimate* or *Underestimate* the number of red marbles, while judges aimed to accurately guess the truth. The number of clicks required to change the color of a marble served as the manual cost for the sender to generate more extreme lies. Specifically, participants either needed to click each marble just once to switch their color (lower `Linear-cost`), or they needed to click each marble an additional time to switch color resulting in the number of required clicks to grow quadratically with n: $1+2+3+...+n = \frac{n(n+1)}{2}$ (higher `Quadratic-cost`). Thus, participants in the `Quadratic-cost` condition needed to exert more effort to produce more extreme lies. If the amount of effort senders committed to trials was consistent between the conditions, then we would expect that senders produce less

bias in their lies when subjected to higher costs in the `Quadratic-cost` condition. Participants were randomly assigned to one of the 2×2 conditions.

Participants were explicitly instructed about the goal of the sender, and the cost to switch marbles (e.g., `Quadratic-cost`: "The more marbles you switch color, the more clicks you'll need to switch each marble.") During `Quadratic-cost` sender trials, a circular progress bar tracked the number of clicks already completed and the number of additional clicks to switch a given marble color. In the `Linear-cost` condition, the circular progress bar was not present. Participants were instructed that the original jar was composed of 50% red and 50% blue marbles. However the marbles were sampled from a beta-binomial distribution $X \sim BetaBinomial(100,3,3)$ (95% of samples fall between 14 and 86), which while still centered at 50, yields more variability than a standard binomial distribution (95% fall between 40 and 60). By increasing the variability of ground-truth, we expected that participants would rely more on their beliefs about cost functions, rather than base-rates, to judge the truth. To avoid counting errors, both the sender and the judge were also explicitly informed of the number of red and blue marbles in the display. Furthermore, to avoid concerns about the positional distribution of marbles serving as a cue to deception, the positions of marbles were shuffled before the senders altered display was shown to the judge.

Participants played against a computer opponent, which allowed us to control for the opponent's behavior. Participants were not explicitly told if their opponent was a computer or a human. The computer opponent's response time, average lying, and inference behavior was held constant across cost function conditions to ensure that any potential variation in participants' judgments of ground-truth was caused by their beliefs about the sender's cost function and not by the computer sender's actual behavior. Specifically, the computer sender lied by taking the truth and adding in the direction of their goal some sampled amount, taken from a Poisson distribution with a mean of 5. As a judge, the computer sampled from the same distribution but subtracted from the participant's message.

Participants played for two practice trials: first as the sender, then as the judge. Then,

participants played for 100 test trials, alternating between sender and judge roles every trial (which role was played first during the test trials was randomized). Throughout the task, participants additionally answered 12 attention check questions related to the trial (two in the practice trials, and ten randomly distributed in the test trials). To prevent participants from relying on learned information about their opponent's behavior, participants did not receive direct feedback about their opponent's decision or the trial's outcome. Instead, they received feedback about the players' cumulative points every five trials, which motivated participants to play the game while only revealing coarse information about their success.

3.1.2 Results

Validating preconditions in lying behavior

We first validated that the condition manipulations worked, and senders chose lies that were driven by their assigned goals and were systematically constrained by the assigned cost function. Senders biased their lies in the same direction as their goal to induce the judge to over- or underestimate (Fig. 3.2). Using linear models with random-effects for subject and item (the true draw), we found that (as expected) senders whose goal was to overestimate inflated their message relative to the truth ($\hat{\beta} = 5.98$, $t(162) = 5.99$, $p < 0.0001$), and those whose goal was to underestimate deflated their message ($\hat{\beta} = -4.46$, $t(132) = -5.50$, $p < 0.0001$). Additionally, although the bias point estimate is systematically larger for senders with the goal to overestimate, there is not a significant difference. We also validated that the cost conditions systematically influenced how senders lied: `Linear-cost` senders introduced more bias into their message relative to the `Quadratic-cost` senders ($\hat{\beta} = 5.15$, $t(202) = 4.81$, $p < 0.0001$), aggregating over goals. These results showed that senders generated lies consistent with their assigned goal and cost function.

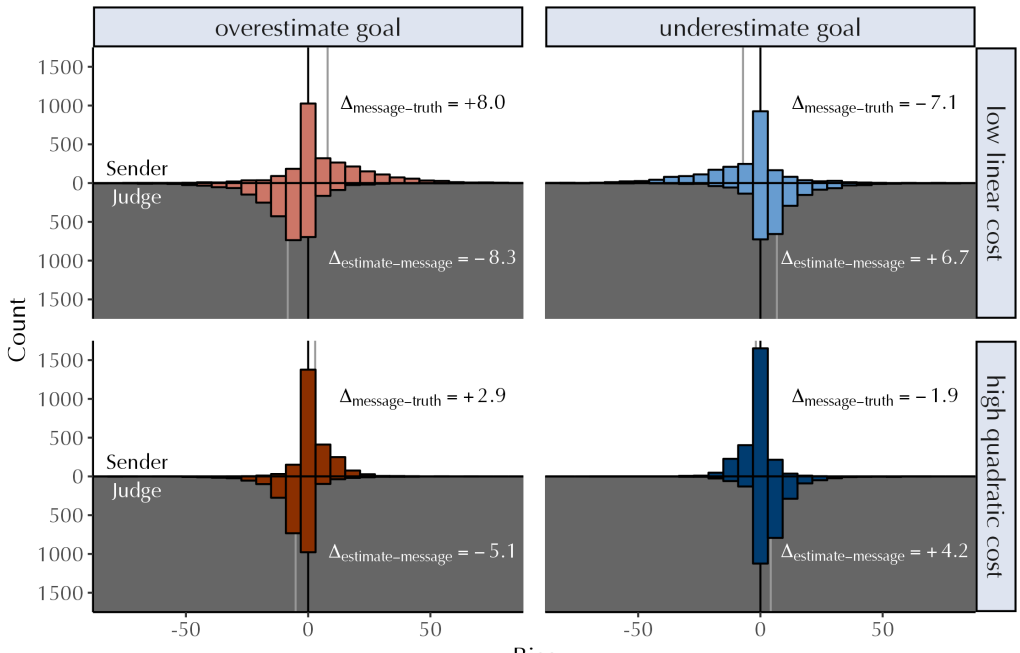


Figure 3.2. Distribution of participant senders' biasing and judges' bias-correcting behavior across each condition (panel columns are goals and rows are cost functions). The direction and distance of the gray line (mean bias) relative to the black line at $bias = 0$ indicates if participants generally inflate (positive bias) or deflate (negative bias) their response and how large the difference is. The top half of each panel (in white) shows how much senders manipulate their message relative to the truth ($\Delta_{message-truth}$). Senders with *Overestimate-goals* (left panels) biased their messages in the positive direction from the truth, and vice versa, senders with *Underestimate-goals* (right panels) biased their messages in the negative direction. Senders with lower `Linear-costs` produced more bias (mean is farther from 0) in their message, than those with higher `Quadratic-costs`. The bottom half of each panel (in gray) shows how much judges adjust their truth estimate relative to the sender's message ($\Delta_{estimate-message}$). Judges bias corrected in the opposite direction of senders' bias – when judges expected senders to have *Overestimate-goals* and bias their message in the positive direction, judges tuned how they bias corrected their estimate in the negative direction. Judges also tuned how they bias corrected to the sender's expected costs – when judges expected senders to have lower `Linear-costs` to lie, they bias-corrected more (mean is farther from 0) in their estimate of the truth. The mean bias for each role and condition is shown in text on the plot.

Do people estimate the truth by considering their beliefs about others' goals and costs?

Judges who apply their beliefs about senders' goals should make bias corrections in the opposite direction of the senders' goal. If the sender wanted the judge to overestimate, then the judge should expect the sender to positively bias their message by adding more red marbles. A judge who expects positive bias in the message should correct for the bias in the negative direction

by estimating that the true number of marbles drawn was fewer than what was reported in the message. As predicted, we found that participants in the *Overestimate* condition bias-corrected in the negative direction by guessing smaller numbers ($\hat{\beta} = -6.84$, $t(148) = -7.31$, $p < 0.0001$). Vice versa, participants in the *Underestimate* condition bias-corrected in the positive direction by guessing larger numbers ($\hat{\beta} = 5.33$, $t(159) = 7.20$, $p < 0.0001$). Once again the bias correction point estimate is systematically larger for receivers in the overestimate goal condition, but there is not a significant difference. These results show that the direction of people's truth inferences are informed by their beliefs about speakers' goals.

Judges that apply their beliefs about senders' cost functions should expect senders with lower cost functions to produce more extreme lies. Therefore, they should make larger magnitude bias corrections. Indeed, judges who believed the sender had a `Linear-cost` debiased their estimate more compared to the `Quadratic-cost` ($\hat{\beta} = -2.96$, $t(202) = -3.53$, $p < 0.001$), aggregating over goals. Figure 3.2 shows that the bias corrections' absolute distance to the intercept is larger for the `Linear-cost` (top panels), compared to the `Quadratic-cost` (bottom panels).

Lastly, when comparing human judge bias correcting to human sender biasing, we do not see any systematic over- or undercorrections. People are not overly nor insufficiently trusting relative to how people would lie in this task. Thus, people broadly seem to calibrate how they correct for bias to how they add bias into their lies. A more thorough investigation into individual differences between individuals' own sender and judge behavior is included in the Supplementary Materials.

We asked whether people can estimate the truth from the content of a lie. We tested the hypothesis that this feat can be achieved without clairvoyance so long as listeners know how speakers are (1) directionally motivated to lie, and (2) cost-constrained in the magnitude of their lies. Our behavioral experiment manipulated the goals and costs of speakers' deception, and showed that participants are sensitive to these factors when lying. Critically, people are also calibrated to the senders' goals and costs when they try to estimate the truth from the content

of the lie, suggesting that in settings where goals and costs are transparent, overt lies may not actually lead to systematic deception.

3.2 Probabilistic Simulations: Consequences for Communication Systems

Our behavioral study showed that listeners can estimate reality from deceptive messages by considering the speaker's motives and costs. The behavioral result suggests that in certain communication channels, senders and judges pass around dishonest messages, yet judges approximately infer the ground truth. This result opens a number of questions about how communication would work in such settings. Consider again recommendation letters. First, if a letter writer predicts readers take away a softened interpretation of writers' claims, then perhaps a writer ought to further amplify their claims about the candidate. If such escalation proceeds unchecked, recommendation letters may become completely decoupled from reality. What are the requirements to keep this process in check, and what properties do we expect of the resulting communication channel?

Second, while some letter writers may embellish their claims, other writers may want to accurately convey their beliefs about a candidate. When there is a mixture of speakers who have varying motivations, listeners may be best served by assuming the speaker is semi-deceptive, semi-cooperative and systematically curb their vigilance about reality accordingly. Under this assumption of listener behavior, dishonest messages will be interpreted nonliterally, and so too will honest messages. Thus, a cooperative speaker, who intends for the listener to extract an accurate interpretation, will fall short if they simply say an honest message. How would cooperative speakers behave in an environment with listeners that expect many deceptive speakers? In the next section, we examine these population-level dynamic using probabilistic modeling.

3.2.1 Model Setup

We consider two interacting agents: senders and judges. Senders observe some ground truth (k) and select a message to say to the judge (k_{say}); thus they are characterized by their utterance distribution $P_S(k_{say} \mid k)$. Judges observe the message from the sender (k_{say}) and produce an estimate of the truth (k_{est}), and so are characterized by their estimation distribution $P_J(k_{est} \mid k_{say})$. The conditional response distributions of senders and judges arise from a decision rule over their expected utilities, calculated from their utility functions. To maintain generality, these utility functions are both defined over $\{k, k_{say}, k_{est}\}$ tuples.

In the basic agent model we consider here, judges want to accurately estimate the truth, so their utility function can be characterized as an L2 loss function on the error of k_{est} relative to k without considering the specific message they received (k_{say}) at all:

$$U_J(k, k_{say}, k_{est}) = -(k_{est} - k)^2 \tag{3.1}$$

While judges have a simple, constant goal to be accurate, it is useful to consider senders with different goals. Broadly, pragmatic senders design a message about the world by considering what beliefs it will instill in the judge.

Deceptive senders (S_D) aspire to mislead the judge by causing them to mis-estimate the truth. This can be captured by utility that scales with the error of the judge's estimate (k_{est}). However, this deceptive sender does not wish to produce messages too far off from reality because of a cost function penalizing increasing falsity in their message. This can be captured by an L2 loss on deviations between message and reality.

$$U_{S_D}(k, k_{say}, k_{est}) = (k_{est} - k) - m(k_{say} - k)^2 \tag{3.2}$$

The parameter m represents a ratio of the deceptive sender's relative desire to induce a biased inference in the listener versus their cost to make messages more discordant from reality.

In contrast, pragmatic cooperative senders (S_C) have goals that align with judges, and thus also want judges to form accurate beliefs about the world, and only consider an L2 loss function on the judge's estimate error:

$$U_{S_C}(k, k_{say}, k_{est}) = -(k_{est} - k)^2 \tag{3.3}$$

This pragmatically-cooperative utility function is notably different from that of a literally honest sender, who would only seek to minimize the deviation of their message from reality ($-(k_{say} - k)^2$) regardless of how that message is understood by the judge. This distinction between considering how a message is interpreted, rather than its literal meaning, is at the heart of modern models of cooperative communication (Frank & Goodman, 2012; Goodman & Frank, 2016), which argue that human language use can be understood in terms of such pragmatic motives. Later, we will see this distinction between pragmatic, and literal, honesty is important when cooperative speakers share a communication channel with liars. Other utility terms may be considered, but for our purposes, this is a minimal set to illustrate the dynamics that emerge in not-entirely-cooperative communication channels.

The decision rule for the sender and the judge are given as the softmax of their expected utilities, where α is the decision noise parameter. Defining these decision rule entails mutual recursion because the sender's utilities depend on the predicted response of the judge

$$P_S(k_{say} \mid k) \propto exp(\alpha \sum_{k_{est}} U_S(k_{est}, k_{say}, k) P_J(k_{est} \mid k_{say})) \tag{3.4}$$

and the judge's utilities depend on inverting the sender's message distribution

$$P_J(k_{est} \mid k_{say}) \propto exp(\alpha \sum_{k} U_J(k_{est}, k_{say}, k) P(k \mid k_{say})) \tag{3.5}$$

where the conditional probability of the truth is given by:

$$P(k \mid k_{say}) = \frac{P_S(k_{say} \mid k)P(k)}{\sum_k P_S(k_{say} \mid k)P(k)} \tag{3.6}$$

We ground out this recursive definition in a level 0 "literal" judge, who interprets the sender's message according to the literal semantics (Goodman & Frank, 2016). The literal judge's estimate of the truth directly matches their received message.

3.2.2 Results

Do messages become increasingly decoupled from reality?

Probabilistic models serve as tools that help explain how emergent properties arise from the interactive dynamics of simple agents. Here, we examine how the properties of the communication channel change as a function of the senders' motives. We first tested how deceptive and cooperative senders adjust their messages to the predicted responses of judges over progressive levels of recursive reasoning. We hypothesize that the communication channel yields one of two potential patterns of stability. (1) Lies and truth inferences are amplified with each level of recursion, becoming increasingly decoupled from reality, and ultimately yield an unstable communication channel. Or (2) lies and truth inferences are checked by constraints of agents' goals, and ultimately converge on a stable, equilibrium state.

The simulation is initiated with the literal judge (Level 0, or L0) who directly estimates k_{est} from k_{say}. Next, a Level N (LN) cooperative or deceptive sender probabilistically decides what k_{say} (conditioned on k) to produce under the assumption that the judge behaves like an $N-1$ thinker. Then, an LN judge probabilistically decides what k_{est} (conditioned on k_{say}) to guess under the assumption that the LN sender behaves rationally. Senders and judges recursively reason in this way on and on. Figure 3.3 shows what senders message conditioned on the truth (green), what judges estimated about the truth conditioned on what message they received (red), and what judges estimated about the truth conditioned on the actual truth (orange).

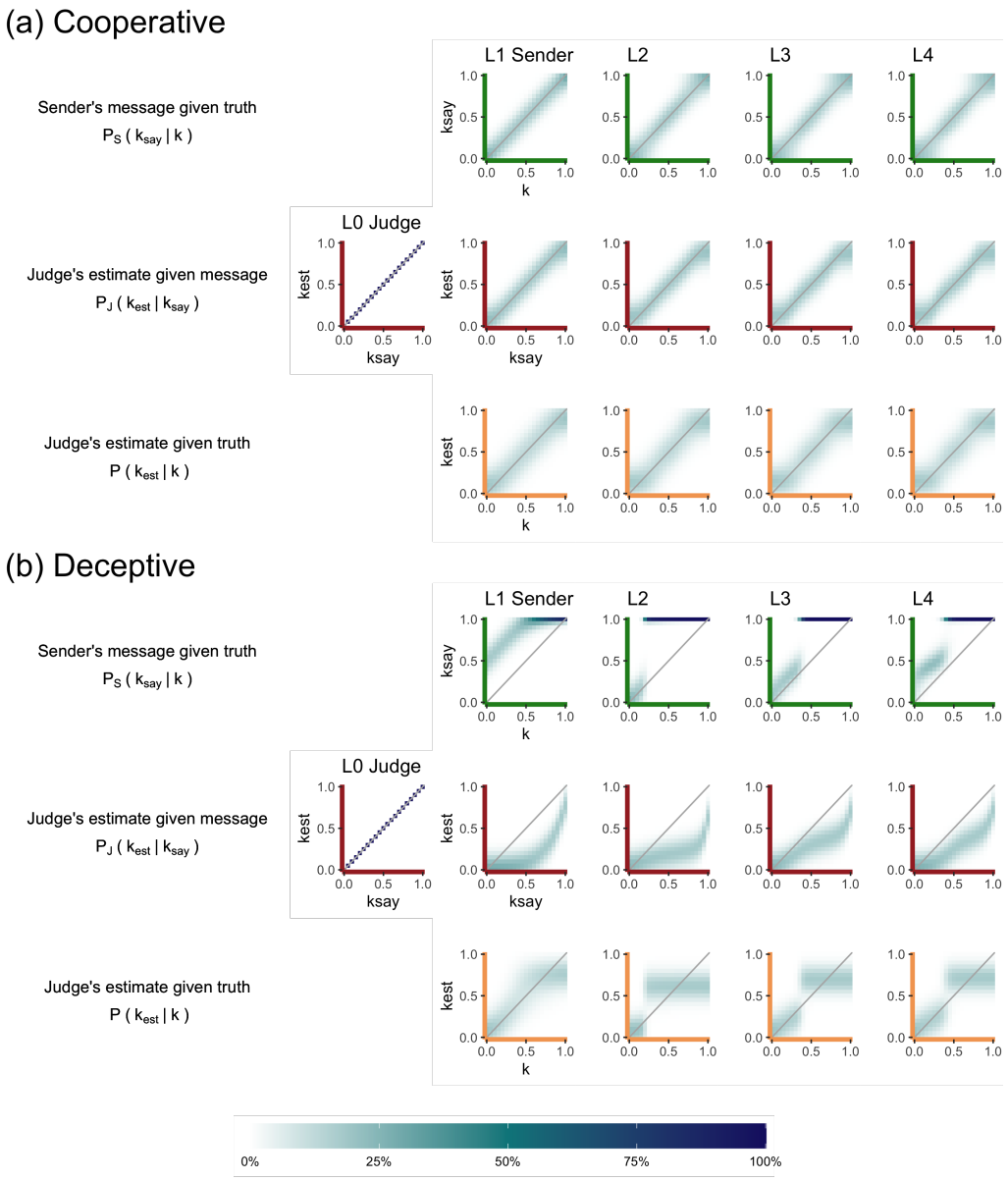


Figure 3.3. Simulated behavior of sender-judge dyads over evolving levels of social reasoning (Levels L0 to L4). The shading reflects the probability the agent performs a behavior given the observation: green plots shows the sender's message conditioned on the truth, or $P_S(k_{say}|k)$; red shows the judge's estimate conditioned on the message they received, or $P_J(k_{est}|k_{say})$; and orange shows the judge's inference about the truth, or $P(k_{est}|k)$. (a) Cooperative senders produce unbiased messages, and judges' inferences about the truth are unbiased and have little noisy. (b) Deceptive senders quickly converge on systematically producing biased messages. Judges' inferences about the truth are noisy but unbiased because they consider the sender's motive to deceive. The ratio of intended bias to message cost (m) is set to 1.

As a basic validation, we show that cooperative senders say honest, unbiased messages, as observed by the shaded region along the identity line ($\Delta_{k_{say}-k} = -6.7 \times 10^{-19} \approx 0$; Figure 3.3). The noise around what cooperative senders say arises from the probabilistic nature of the agents' decision rules. In turn, judges who expect the sender to be cooperative tend to interpret messages literally ($\Delta_{k_{est}-k_{say}} \approx 0$) and their resulting truth inferences are unbiased ($\Delta_{k_{est}-k} \approx 0$).

In contrast, deceptive senders say dishonest, biased messages ($\Delta_{k_{say}-k} = 0.28$). Critically judges who expect the sender to be deceptive correct for that bias in their interpretation of messages ($\Delta_{k_{est}-k_{say}} = -0.24$). Ultimately judges' resulting truth inferences are substantially less biased ($\Delta_{k_{est}-k} = 0.006$) than the messages and estimate, even if their estimates are noisier ($R^2 = 0.61$) than those of judges paired with cooperative senders ($R^2 = 0.74$). Importantly, even with greater recursion levels, messages converge to a stable equilibrium state, rather than becoming ever-increasingly decoupled from reality.

How do senders' motives determine equilibrium form?

We propose that the sender's relative motive to have the judge mis-accurately infer the truth to the cost function to produce more extreme lies, or m, influences the sender-judge equilibrium state. We simulated 111 unique deceptive senders and varied the value of m between $\frac{1}{3}$ and 10. To ensure that agent pairs converged on an equilibrium state, we examined pairings at the (arbitrarily chosen) 20^{th} level of recursive reasoning. As m increases, we find that, at equilibrium, deceptive senders trend towards being more dishonest and thus produce a larger bias (Fig. 3.4). So too does the judge's bias correction in the opposite direction. Then, critically, we ask how m impacts judges' inferences about the truth ($k_{est} - k$) at equilibrium by testing how (1) accurate and (2) precise are judges' estimate of the truth relative to the ground truth.

We may expect accurate, or unbiased, truth inferences, in which case judges at equilibrium can perfectly correct for the bias introduced by senders. Alternatively, judges may *under*correct for senders' bias, so senders "win" in the long run because they succeed at causing the judge to overestimate. Or judges may *over*correct for senders' bias, so senders incidentally self-sabotage

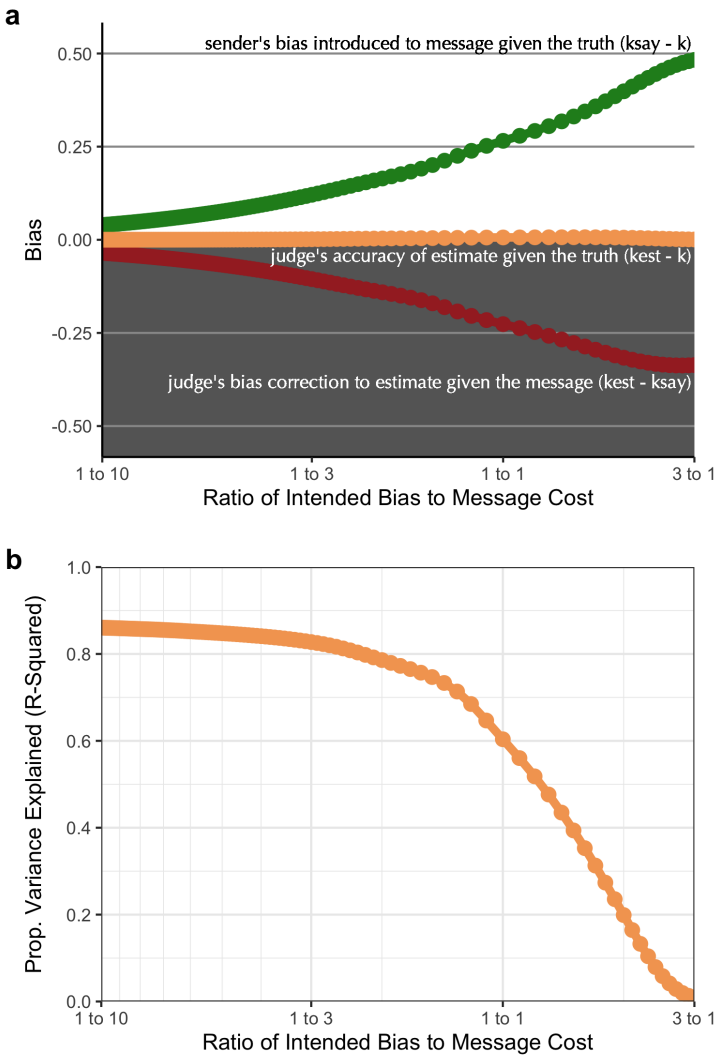


Figure 3.4. Model's predicted bias and precision as a function of the ratio m. The x-axis is log scaled, with lower values representing when message cost dominates intended bias. (a) Senders' biases in their message relative to the truth (green; $k_{say} - k$) and judges' bias correction in their estimate relative to the message (red; $k_{est} - k_{say}$) both increase absolutely at higher ratios m, when intended bias dominates message cost (get further from 0). Regardless, the accuracy of judges' truth inference relative to the ground truth (orange; $k_{est} - k$) stabilizes at 0, implying that the truth gets unbiasedly conveyed to the listener in these communication systems even when messages are lies. (b) The proportion of variance explained (R^2) by the truth in judges' truth inferences decreases with higher ratios, meaning that there is less precision in communication systems in which speakers face relatively lower costs to lie.

by leading judges to underestimate (counter to senders' goals). Each of these predictions appraise the success of the communication channel: do judges accurately extract the truth from

communication, or do deceptive senders succeed at distorting judges' beliefs?

The model finds that judges' truth inferences are unbiased across all ratios m. Senders ultimately fail to distort judges' beliefs. This surprising finding calls into question the benefit to deceivers to lie when their motives are broadly suspected. A related configuration of deceptive communication channels is that while they may not bias the judges' inferences, the process may result in increased imprecision. For example, even if truth inferences were unbiased across repeated interactions, judges could still be inaccurate for most individual interactions.

How does precision in judges' truth inferences change as a function of the ratio m? We measured R^2, or the proportion of estimates' variance explained by the truth. Larger R^2 implies that judges' truth inferences are more consistent with the ground truth, while smaller implies that they are more distributed. The model finds that as the ratio of m increases, R^2 decreases to 0. In other words, truth inferences are more distributed when senders face a relatively lower cost to lie.

In sum, we found that deceptive senders' relative intended bias to distort judges' beliefs versus their message cost to produce larger lies drives the form of equilibria. In particular, we found that bias for both senders' messages and judges' inferences increases as intended bias increasingly dominates cost. Furthermore, we found that while accuracy in judges' truth inference is unbiased, precision decreases as intended bias increasingly dominates cost. Thus in communication channels where speakers face fewer costs to lying and judges suspect this, messages become more divorced from reality; nonetheless people extract accurate, albeit noisier, information about the truth.

How do judges' bias corrections influence speakers with different motivations?

Populations are composed of agents with varying motivations. While some speakers may be deceptive, more often speakers aim to be cooperative (Grice, 1975). How may speakers indirectly influence one another via socially reasoning about judges' behavior? In our model, speakers only directly interact with judges, not other speakers. This design allows us to isolate how speakers indirectly influence other speakers by way of judges' beliefs and actions. Specif-

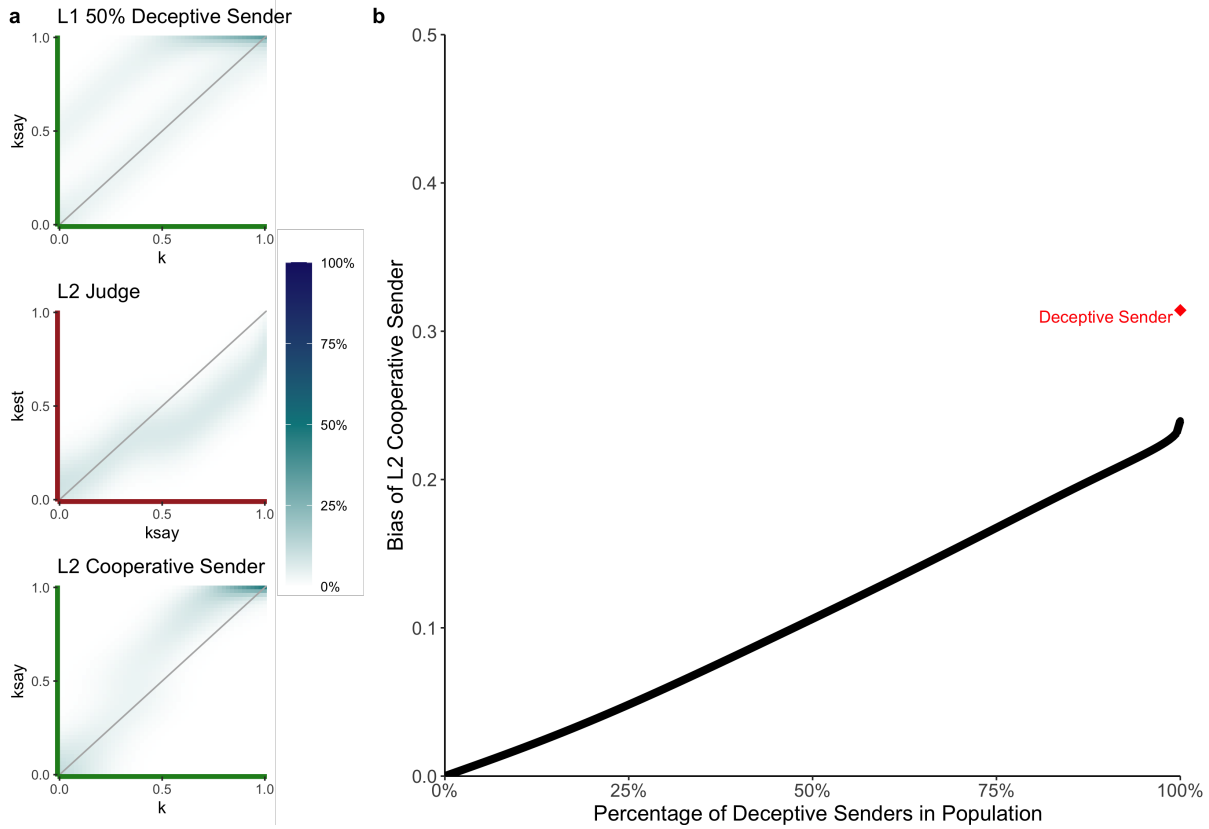


Figure 3.5. Simulated behavior of a cooperative sender in a population with deceptive speakers. (a) A cooperative sender assumes the judge believes the population of senders is 50% deceptive and 50% cooperative. The top panel shows the bias of a mixture of messages from cooperative (on the identity line) and deceptive (off the identity line) L1 senders. The middle panel shows the L1 judges' bias correction for this mixture. The bottom panel shows how a cooperative L2 sender would bias their message (instead of being honest) to cater to the judge's bias correction. (b) The bias of L2 cooperative senders' messages increases with higher percentages of deceptive senders in the population. At 100% deceptive populations, L2 cooperative senders bias their messages less than L2 deceptive senders (red rhombus).

ically, we explore how cooperative speakers produce messages when they think the judge is correcting for deceptive speakers in the population?

We examined the behavior of a cooperative L2 sender under the L1 judge's assumption about a mixed population of deceptive and cooperative senders. We varied the proportion of deceivers in the population, which scales how much bias the L1 judge assumes in the message and therefore how much they bias correct. The L1 judge's expected estimate of the truth gets fed to an L2 cooperative sender. Critically, to help the judge accurately infer the truth, the L2

cooperative sender adjusts their behavior to produce a *dishonest* message. Figure 3.5a shows an example simulation when the population is 50% deceptive.

We found that as the proportion of deceivers in the population increased, the L2 cooperative sender produced more bias in their messages (Fig. 3.5b). At the upper limit, when the cooperative sender expects the judge to believe the population is composed of 100% deceivers, the cooperative sender produces a bias of 0.24, which is still less than a deceptive sender with the same belief about the population, who produces a bias of 0.31. While the cooperative sender still biases their message, they do not do so to the extent of their deceptive counterpart.

3.3 Discussion

Detecting lies is often portrayed as a categorization process – is a message true or false? However, in many real world situations, people go beyond simply categorizing to make richer inferences – what is the actual ground truth? We focus on one goal in deception – that a speaker wants the judge to mis-estimate the truth (e.g., my candidate is the best fit for your position), while a judge's goal is to infer the truth (e.g., your candidate is an okay fit). Altogether, our studies show that a rational theory-of-mind framework explains how people may infer the truth from literally false messages. Behaviorally, we show that people can and do generate inferences about the truth from suspected lies, and they tune these inferences to their beliefs about senders' motives and costs. When speakers and listeners have veridical representations of each others' adversarial motivations and costs, the result is a state in which speakers say literally false messages but listeners nonetheless extract the truth. Probabilistic modeling shows speakers do not ratchet up to produce more and more extreme lies with increasing levels of theory-of-mind reasoning; instead, the communication channel stabilizes to an equilibrium state. For broad classes of systems, social reasoning about others predicts how accurately and precisely communication channels (e.g., letters of recommendation) transmit information about the ground truth. For individuals within systems, speakers' different motivations indirectly affect what

others say, so that even cooperative speakers should be dishonest when they suspect listeners are correcting for deceptive speakers (e.g., cooperative recommenders should embellish what they say to accurately convey their belief to the reader).

While rational and recursive reasoning form the framework to explain how people infer truth from lies, two critical components serve as a precondition for such a system to get off the ground, that: (1) listeners know of speakers' directional deception goals and (2) bigger lies are more costly. Throughout this paper, we highlighted letters of recommendation as a communication system which, in equilibrium, messages are biased – recommenders inflate how positively they write about their candidate – yet the transmission is unbiased – readers extract accurate beliefs about the candidate. We speculate that these critical components coexist within many real-world communication systems, and thus our unified framework can explain idiosyncratic behaviors in communication systems that have not been linked previously. For example, when communicating your preferred political candidate via voting in run-off elections, voters can be honest by selecting their favorite candidate, or "strategic" by misrepresenting their preference in earlier rounds (Piketty, 2000). Both voting methods converge to different equilibria states – one is honest and one is dishonest, yet both result in the overall preferred candidate being elected. In essence, honesty and dishonesty are equally serviceable solutions to transmitting information. Then, there is puffery in marketing, which may not be perceived as false advertising because listeners make the adjustment to how they interpret the message (Stern & Callister, 2020). Even in communication systems that are not standardly thought of as deceptive, these principles may be applied to understand why populations form norms to produce nonliteral messages, such as in commonplace hyperbole in everyday language (Bennett, 2015).

Until now, we have treated deceivers and cooperators as distinct agents. We defined cooperators as sharing the same goal as judges: to induce the judge to form an accurate belief about the world. Formally, we characterized deceivers as senders who have weighted incentives to induce a (mis-)belief in the listener, but face weighted costs to saying more extreme lies. Cooperators can be mapped onto this formalism as well. Cooperators want to reduce the error in

the accuracy of judge's estimates of the truth, juxtaposing deceivers that want to induce error. In this paper, we highlighted a cooperator that places zero weight on how they deviate their message from reality, although in principle cooperative senders may prefer to be honest, as deceivers do. Thus, our framework presents a unifying factor between cooperators and deceivers in their motive to influence the listener's beliefs.

The cost of lying also plays a critical role in driving how communication appears in these distorted communication systems. Our simulations showed that as cost functions decrease, messages are more dishonest. Cost drives communication systems to approach an equilibrium state that preserves accurate inferences about the truth, even when speakers are deceptive. Formally, the emergence of equilibrium depends on a crossover effect between the linear incentive to distort listeners' belief versus the quadratic cost to producing larger lies. When this crossover effect is no longer valid – in our formalization, this happens when the incentive to distort listeners' belief far surpasses the cost – the communication system takes on a ratcheting effect, in which lies become increasingly extreme. When lies are so extreme that they become decoupled from reality, listeners no longer extract signal from the message, so accurate inferences to the truth end up as an incidental byproduct of listeners randomly guessing what the truth could be. This "runaway lies" effect points to the value of intensifying liars' costs to producing more extreme lies. Whether lying costs are increased via interventions targeting individuals' cognitive load or reputational risks, or by improving detection algorithms, listeners would gather more signal from lies and more precisely infer the truth.

Now that we have characterized communication systems that advantage listeners, conversely we can better understand when listeners' inferences go awry. While people may be aware of broad goals within a given communication systems (e.g., letter writers generally want to promote their candidate), they may not be fine-tuned to individuals' motives and costs. For example, in general people would not suspect that a letter is downright fabricated, assuming that most people face higher costs to drastic deceptions. Therefore, individual fabricators benefit from readers who under-correct the bias by assuming that the letter is embellished, but not that it

is fabricated. Meta-reasoning deceivers may even actively conceal their motives and costs. Of course, speakers who fake a cooperative intention build trust with the listener and are the most successful deceivers. But an even richer (untested) prediction from this work is that speakers, even when transparent about their deceptive intent, can conceal how *strong* their intention is to deceive to *gradedly* dupe their listener.

In conclusion, people's intuitive theory-of-mind reasoning – and not necessarily the assumption that others are cooperative – allows listeners to infer the truth from literally false messages, so long as they are equipped with sufficient knowledge about the speakers' goals. Taking a first principles approach to agents' goals, costs, and actions, we bridged individual listeners to broad classes of distorted communication systems, to characterize how they both systematically transmit and interpret information. Lastly, these results call into question the traditional depiction of people as naïve lie detectors, and instead support a nuanced depiction of people as robust lie interpreters.

3.4 Acknowledgments

This material is based upon work supported by the NSF Graduate Research Fellowship under Grant No. DGE-1650112 to LAO. We would like to thank Frank Mollica, Judy Fan, Robert Hawkins, Lindsey Powell, Mike McCullough, Holly Huey, and Aarthi Popat. Lastly, we thank Chaz Firestone for inspiring this work with his Tweets.

Data and code for the experiment and analysis are available at `https://github.com/la-oey/WhatIsReality`

Chapter 3 is a reprint of material that has recently been submitted to *Computational Brain & Behavior* and appears as Oey, L. A., & Vul, E. (under review). Accurate approximations about the truth from literally false messages. An early version of the project was published as a conference proceeding: Oey, L. A., & Vul, E. (2022). Inferring truth from lies. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the*

Cognitive Science Society (pp. 1469-1475). The dissertation author was the primary investigator and author of both these paper.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115–1153.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires, and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1–10.

Bennett, J. (2015). OMG! the hyperbole of internet-speak. *The New York Times*. https://www.nytimes.com/2015/11/29/fashion/death-by-internet-hyperbole-literally-dying-over-this-column.html

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234.

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, *134*(4), 477–492.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Debey, E., De Houwer, J., & Verschuere, B. (2014). Lying relies on the truth. *Cognition*, *132*(3), 324–334.

Feiler, D. C., Tong, J. D., & Larrick, R. P. (2013). Biased judgment in censored environments. *Management Science*, *59*(3), 573–591.

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

Franke, M., Dulcinati, G., & Pouscoulous, N. (2020). Strategies of deception: Underinformativity, uninformativity, and lies — misleading with different kinds of implicature. *Topics in Cognitive Science*, *12*(2), 583–607.

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological Science*, *20*(3), 393–398.

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics Vol. 3: Speech Acts* (pp. 64–75). Academic Press.

Gweon, H., Pelton, H., Konopka, J. A., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when teachers are under-informative. *Cognition*, *132*(3), 335–341.

Hayes, B. K., Banner, S., Forrester, S., & Navarro, D. J. (2019). Selective sampling and inductive inference: Drawing inferences based on observed and missing evidence. *Cognitive Psychology*, *113*, 101221.

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science*, *24*(5), 379–385.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(10), 589–604.

Leach, A.-M., Talwar, V., Lee, K., Bala, N., & Lindsay, R. C. L. (2004). "intuitive" lie detection of children's deception by law enforcement officials and university students. *Law and Human Behavior*, *26*(6), 661–685.

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communication Monographs*, *66*(2), 125–144.

Lewis, D. (1969). *Convention: A Philosophical Study*. John Wiley & Sons.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, *152*(2), 346–362.

Piketty, T. (2000). Voting as communicating. *Review of Economic Studies*, *67*(1), 169–191.

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of sciences*, *105*(3), 833–838.

Ransom, K., Voorspoels, W., Navarro, D. J., & Perfors, A. (2019). Where the truth lies: How sampling implications drive deception without lying. *PsyArXiv*.

Schelling, T. C. (1960). *The Strategy of Conflict*. Harvard University.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Stern, L. A., & Callister, M. (2020). Exploring variations of hyperbole and puffery in advertising. *Journal of Current Issues and Research in Advertising*, *41*(1), 71–87.

ten Brinke, L., Vohs, K. D., & Carney, D. R. (2016). Can ordinary people detect deception after all? *Trends in Cognitive Sciences*, *20*(8), 579–588.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691.

van't Veer, A. E., Stel, M., & van Beest, I. (2014). Limited capacity to lie: Cognitive load interferes with being dishonest. *Judgment and Decision Making*, *9*(3), 199–206.

Walczyk, J. J., Harris, L. L., Duck, T. K., & Mulay, D. (2014). A social-cognitive framework for understanding serious lies: Activation-decision-construction-action theory. *New Ideas in Psychology*, *34*, 22–36.

Chapter 4

Conclusion

4.1 Individuals, dyads, and collectives

People are capable of impressive strategic feats. In this dissertation, I propose that people can skirt the boundaries between cooperation and defection, challenging traditional binaries set by precedence (e.g., the Prisoner's Dilemma). People are not simply cooperating or defecting—by deploying recursive social reasoning, they can do both. They can strategically select how to defect, such that they appear to cooperate. More broadly, social cognitive mechanisms allow people to jointly pursue their goals and maneuver their local social environment. From these strategic and adversarial behaviors, collective behaviors and norms may emerge.

Each chapter highlights instances of how individuals make strategic decisions when embedded within their social environment. Chapter 1 applies data scientific techniques to analyze social decision making in networks. In particular, it explores the cognitive scientific community as a case study for how individuals' parallel decisions—to co-authorship with others and study select research topics—coalesce to drive the collective's integrative behaviors. Chapters 2 and 3 apply computational models and behavioral studies to quantify the flexibility of human deceptive communication. Chapter 2 tests how individuals strategically flout cooperation by crafting lies that thwart audiences' detection within dyadic interactions. Chapter 3 delves into downstream consequences of strategic lying, namely how do audiences interpret suspected lies to infer the truth, and as a result, what emergent properties arise in the collective's communication

systems. These chapters interweave strategic behavior across social group levels, collectives to dyads to individuals and their various interactions. Although there is more theory to be developed, more empirical evidence to be gathered of human strategic behaviors, and more links to be drawn across social levels, this dissertation paves the way toward a unifying theory of individual and collective social intelligence.

4.2 Partial observability of intentions

One presumption in these chapters is that there is full observability of adversarial intentions. Listeners broadly believe that speakers are motivated to lie, and speakers broadly believe that listeners are vigilant. This full observability of intentions exists in real world situations too. For example, dating app users are a priori aware that other people's profiles may feature outdated or filtered photos, list fake heights or ages, over-embellish how active they are in outdoor sports, diminish how often they factor their astrology sign into their everyday decisions, etc. For each dating app user, the goal to represent one's "best self" is transparent. When each user is mutually aware of these representational goals, a cultural norm arises. It becomes permissible for self-presentations to deviate from the truth.

In other scenarios (different from dating apps), lying well requires concealing one's true adversarial goals and intentions. These lies occur in social environments where people are a priori expected to behave cooperatively. Hiding intentions requires more than simply omitting one's adversarial intentions. At a meta level, cooperative people signal their cooperative intent, e.g., by appearing uncalculating (Jordan et al., 2016). My work thus far shows that people are socially adaptive lie detectors, counter to previous literature that broadly paints people as gullible. Perhaps this dichotomy could be explained by audiences' prior beliefs or deceivers' signaling of intent. If so, encouraging people to pay attention to intent, in addition to whether something is true (Pennycook et al., 2021), could further support people's ability to discern real from fake news on social media, for example.

4.3 Strategic reasoning for socially intelligent machines

An important value of developing computational models of human behavior, as I do in the works described here, is that we can transfer what we know about human systems to engineer better AI systems. In particular, by understanding the representational structure of socially intelligent humans, we can build structured human-like machines that better understand and help people (Lake et al., 2017). Most recently, systems like ChatGPT, which have been trained on more text and webpages than any single human can read, offer to assist humans in a human-like, conversational manner. However, in its current form, ChatGPT suffers from key theory-of-(human)-mind failures. For example, ChatGPT has a tendency to confidently produce plausible-sounding but downright false knowledge. If people socially engaged with ChatGPT as they would with humans, the apparent confidence could easily mislead people. Thus, in AI's current form, there is a need for people (and AI) to strategically reason about the other's mind and potential miscommunications that arise. AI has a responsibility to recognize faulty queries provided by humans and teach people to engage with the AI better. And people have the responsibility to attribute an "AI" mind to these systems, rather than over-assuming the social competencies of AI.

What are the pathways by which our broader psychological understanding of human strategic reasoning can improve AI systems? Strategic games, like chess (Campbell et al., 2002), Go (Silver et al., 2017), poker (Moravčík et al., 2017), and more recently, Diplomacy (Meta Fundamental AI Research Diplomacy Team et al., 2022), have a long history within AI of being used to evaluate machine intelligence. However, unlike these game environments, everyday social environments are highly variable, and people can flexibly adapt how they interact. By creating better models of human social and collective behavior across social environments, we can develop more diverse benchmarks to evaluate how well current AI systems emulate human social understanding (Burnell et al., 2023).

References

Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., Rutar, D., Cheke, L. G., Sohl-Dickstein, J., Mitchell, M., Kiela, D., Shanahan, M., Voorhees, E. M., Cohn, A. G., Leibo, J. Z., & Hernandez-Orallo, J. (2023). Rethink reporting of evaluation results in AI. *Science*, *380*(6641), 136–138.

Campbell, M., Hoane Jr, A. J., & Hsu, F.-H. (2002). Deep Blue. *Artificial Intelligence*, *134*(1-2), 57–83.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Meta Fundamental AI Research Diplomacy Team, Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., … Zijlstra, M. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, *378*(6624), 1067–1074.

Moravčík, M., Schmid, M., Burch, N., Lisỳ, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., & Bowling, M. (2017). Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, *356*(6337), 508–513.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Dreissche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359.

Appendix A

Supplementary Materials to Designing and detecting lies by reasoning about other agents

A.1 Model

A.1.1 Probabilistic Models

Our models of lying and lie detection assume the probability of an action follows a Luce choice rule based on the expected utility of the action relative to alternative actions, with softmax parameter α:

$$P(\theta) = \underset{\theta}{softmax}(\mathrm{EV}[\theta]) = \frac{\exp(\alpha \mathrm{EV}[\theta])}{\sum_{\theta'} \exp(\alpha \mathrm{EV}[\theta'])} \tag{A.1}$$

The receiver chooses to call BS following a softmax rule weighting of the expected value of calling BS or accepting k^*:

$$P_R(BS \mid k^*) = \underset{BS}{softmax}(\mathrm{EV}_R[BS \mid k^*]) \tag{A.2}$$

The expected value of calling BS is obtained by marginalizing over the possible true ks:

$$\mathrm{EV}_R[BS \mid k^*] = \sum_k U_R(BS; k^*, k) P(k = k^* \mid k^*) \tag{A.3}$$

where $U_R(BS; k^*, k)$ is the payoff for the receiver associated with calling BS or not, given k^* and whether or not it corresponds to the true k. The probability of a reported k^* being true is given by

$$P(k = k^* \mid k^*) = \frac{\sum_k P(k) P_S(k^* \mid k) P(k = k^* \mid k, k^*)}{\sum_k P(k) P_S(k^* \mid k)} \tag{A.4}$$

relying on the prior probability of k (here: $P(k) = Binom(k \mid p, n)$ where p is the probability of success on a single trial and n is 10, the total number of trials), and the probability that the sender would produce a given k^* in response to seeing a particular k, $P_S(k^* \mid k)$.

The sender chooses k^* based on a softmax weighting of the expected value of different reports:

$$P_S(k^* \mid k) = \underset{k^*}{softmax}(\mathrm{EV}_S[k^* \mid k]) \tag{A.5}$$

with the expected values given by

$$\mathrm{EV}_S[k^* \mid k] = \sum_{BS} U_S(k^* \mid BS, k^* = k) P_R(BS \mid k^*) \tag{A.6}$$

where $U_S(k^* \mid BS, k^* = k)$ is the payoff for the sender when reporting k^* given whether BS was called and whether k^* was a lie. Calculating these expected values requires that the sender consider the probability the receiver would call BS for a particular k^*, $P_R(BS \mid k^*)$.

Recursive Theory-of-Mind Model – Senders and Receivers

The receiver's likelihood that they call BS $P_R(BS|k^*)$ (A.2) feeds into the sender's expected value for generating a given report $\mathrm{EV}_S[k^*|k]$ (A.6). Hence, in the recursive theory-of-mind (ToM) model, the choices of what k^* to say and whether to call BS forms a single recursive process. Steps reflect the likelihood of players making decisions, terminating at that level of recursion. At step $\omega = 0$, a base receiver calls BS with 50% probability uniformly across all reported k^*. Then, a 0-level sender reports k^* that is the best response to the base receiver's uniform strategy, following from (A.5). A 1-level receiver then responds by calling BS that is the best response the 0-level sender, following from (A.2), etc.

To prevent infinite recursion, we assume players judge their opponent to be a Poisson-weighted recursive step reasoner (Camerer et al., 2004). This assumption reflects the intuition that we believe others have some limitations in their willing to reason through all possible steps; instead, many players reason through just a few steps, while a smaller subset of players reason deeper. This amounts to the model treating the likelihood of the opponent's possible predicted action as a weighted average of all potential ω-steps of reasoning, where steps follow a Poisson distribution with mean and variance ℓ. The player, after evaluating their opponent's potential behavior, chooses the best response. Thus, the overall predictions of the recursive ToM agents' behaviors are the players' best response to the weighted likelihoods for the opponent's potential action summed over all levels of recursion.

Thus far, the ideal observer senders and receivers we consider assume that players' subjective value for different outcomes is determined solely by the game's payoffs. However, it has been widely shown that people are averse to lying (Gneezy et al., 2013) or are generally averse to risk (e.g., Arrow, 1965; Pratt, 1964). Whether this aversion arises from moral concerns (Goldstone & Chin, 1993; Mazar et al., 2008) or cognitive demands (Capraro et al., 2019; van't Veer et al., 2014; Verschuere et al., 2018; Vrij et al., 2006), such an aversion may be formalized in terms of a penalty to the sender's subjective utility when they report something untrue. We fit such a lying aversion parameter, as a constant penalty to utility (η), by adding it to the sender's utility function when they lie:

$$U_S(k^* \mid BS, k) - \eta(k^* \neq k) \tag{A.7}$$

The receiver is similarly averse to calling BS, which is predicted by a general risk aversion:

$$U_R(BS; k^*, k) - \eta(BS) \tag{A.8}$$

We applied this penalty only as the player was choosing a best response to their opponent's marginalized behaviors, and not at each level of recursion, to simplify the interpretation.

0^{th} Order Theory-of-Mind Model – Senders and Receivers

In the 0^{th} Order ToM model, the model makes assumptions about how an opponent that lacks ToM might behave. Here, a 0^{th} Order ToM sender assumes that their opponent behaves randomly—the opponent receiver calls BS with 50% probability uniformly across all reported k^*. In a similar vein, a 0^{th} Order ToM receiver assumes that their opponent sender is equally likely to report any k^* value regardless of their true k.

A.1.2 Heuristic Models

Equal Intrinsic Aversion Heuristic Model – Senders

The Equal Intrinsic Aversion Heuristic model is an account of lying behavior in which all senders lie by some amount over the truth that is constant for all k. We tuned a Poisson-distributed free parameter $\lambda_{\forall}$ as the difference between reported k^* and true k (here, given as Δ). Δ is $k^* - k$ when senders get points for red, and it is $k - k^*$ when senders get points for blue.

$$P_S(k^* \mid k) = \frac{{\lambda_{\forall}}^{\Delta} e^{-\lambda_{\forall}}}{\Delta!} \tag{A.9}$$

Since reports in the lying game are bounded between 0 and 10, you can only inflate your report by so much, depending on k. For example, when you are motivated to report higher values and $k = 10$, you cannot lie any higher. To deal with this when fitting the model, we normalize Δ to the size of its upper bound for a given k.

Additionally, negative numbers in any Poisson distribution have 0% probability of occurring. So lies in the opposite direction (e.g., reporting 3 when $k = 5$ in the red utility condition) would be catastrophic for fitting a Poisson distribution. To deal with this, we fit a lapse rate that reports uniformly between 0 and 10 with some probability.

Unequal Intrinsic Aversion Heuristic Model – Senders

Work in the literature has suggested that most people are honest, and many lies are told by a few prolific liars (Serota et al., 2010). So, building off the above heuristic account, we consider that some people tell the truth and some people lie in the same fashion as the Equal Intrinsic Aversion model. Some proportion of reports are always the truth and some proportion are a lie by some constant over the truth. We assume that with some probability t people tell the truth. If they do not tell the truth, the probability that they say k^* given k follows Equation (A.9), multiplied by the probability they are not telling the truth. As in the previous model, we fit a free parameter $\lambda_{\exists}$ which tunes Δ, the distance between k^* and k.

$$P_S(k^* \mid k) = \begin{cases} k^* = k & t \\ k^* \neq k & (1-t)\frac{\lambda_\exists^{\Delta} e^{-\lambda_\exists}}{\Delta!} \end{cases} \tag{A.10}$$

Null Hypothesis Significance Testing – Receivers

Now we can consider how a naïve receiver might behave if they are not provided information about the payoff structure. Here, receivers may judge what seems like a lie, while only having access to the base rate information. The best such a receiver can do is to perform null hypothesis testing Φ to determine what reports are unlikely to occur by chance alone. k is largely drawn from a binomial distribution in our empirical data (see A.2.2 Sampling Procedure), so this model is akin to a two-sided binomial test for statistical significance. We allowed p to be a free parameter, and so the midpoint of the significance test varied as a function of the condition.

$$P(BS \mid k^*) = \begin{cases} k^* < E[k] & \Phi(k^* \mid k, p) \\ k^* > E[k] & 1 - \Phi(k^* \mid k, p) \end{cases} \tag{A.11}$$

A.2 Experiment 1

A.2.1 Participants

A total of 228 participants were recruited from the undergraduate population at the University of California, San Diego. The attention check criteria asked that participants respond 75% or greater correct (within an absolute error of 1) on 12 comprehension questions (e.g., as the sender, "How many red marbles did you actually draw?"; as the receiver, "How many red marbles did the other player report?") distributed throughout the task. After exclusion, 212 participants were included in our final data set. Participants were randomly assigned approximately evenly across conditions ($n_{red,20\%} = 35$; $n_{red,50\%} = 36$; $n_{red,80\%} = 37$; $n_{blue,20\%} = 35$; $n_{blue,50\%} = 30$; $n_{blue,80\%} = 39$).

A.2.2 Procedure

Participants played against a computer in the lying game across 100 trials. In each round of the game, both players were presented with a box containing red and blue marbles. The distribution of marbles varied by the prior probability condition: the probability of drawing a red marble was either 20%, 50%, or 80%. Participants gathered this information by visually observing the distribution of marbles in the box. Participants alternated between each trial playing as the sender who could generate lies and the receiver who could detect lies. Participants were randomly assigned to either starting as the sender or as the receiver. Senders either got points for red marbles and receivers received points for blue marbles, or the reverse.

Sender. On each trial, the sender randomly sampled 10 marbles from the box, of which k were red. They then reported how many red marbles k^* they wanted their opponent to *think* they sampled. The sender could choose to tell the truth and report the true number of sampled red marbles, or lie and report any false number between 0 and 10.

Receiver. The true sample was hidden from the receiver. The receiver saw how many red marbles the sender reported, and then chose to accept the reported value or reject it as a lie (call BS).

Payoffs

In the red payoff condition, if the receiver accepted the reported red marbles sampled k^*, the sender got points for the number of reported red marbles, and the receiver got points for the corresponding number of blue marbles. If the receiver rejected but the sender told the truth, the sender got points for the reported red marbles as before, but the receiver got a -5 penalty to the blue marble points. However, if the receiver rejected and the sender lied, the sender was penalized by getting -5 points (regardless of how many were reported) and the receiver was rewarded with $+5$. In the blue payoff condition, all point values were reversed: the sender got points for blue marbles, and the receiver, for red marbles.

Table A.1 shows the players' utility (the point differential) for the *red* payoff condition. The points depended on the sender's reported value k^*, the sender's decision to (a) tell the truth

or (b) lie, and the receiver's decision to (A) accept or (B) reject, or call BS:

		Sender	
		a) Truth $k = k^*$	b) Lie $k \neq k^*$
Receiver	A) Accept	$2k^* - 10$ / $10 - 2k^*$	$2k^* - 10$ / $10 - 2k^*$
	B) Reject	$2k^* - 5$ / $5 - 2k^*$	-10 / 10

Table A.1. Payoff matrix for the game and utilities in the model: the point differential (player - opponent) in the *red* payoff condition.

Sampling Procedure and Computer's Generative Process

Over the course of the 100 trials, every participant saw one instance each of true $k = \{0 : 10\}$ as the sender and one instance each of reported $k^* = \{0 : 10\}$ as the receiver, for a total of 22 predetermined trials that were randomly interleaved among the other trials. In the other 78 trials, marbles were sampled from a binomial distribution $k \sim B(10, p)$. The predetermined trials were included in order to guarantee the inclusion of all sampling values, including those that would occur only rarely under a binomial distribution.

Like the human senders, the computer sender sampled marbles from a binomial distribution $k \sim B(10, p)$. To determine what to report, the computer sender used a mixed strategy: 20% of the time, the computer reported a random sample from a uniform distribution $k^* \sim U(0, 10)$; 80% of the time, the computer independently sampled from an (unseen) binomial distribu-

tion $k' \sim B(10, p)$, and reported the higher number between the true and the intrinsic sample $k^* = MAX(k, k')$.

The computer receiver called BS following a logistic curve with k^* as the input and that shifted with the base rate. It is important to note that after the practice trials, participants never observed the computer receiver's BS calling decision (see A.2.2 Limited Feedback).

Limited Feedback

To ensure that participants were making inferences about how the opposing agent would behave, we designed the task so that participants were not simply using feedback to learn about their opponent over the course of the trials. Thus, participants did not receive feedback about player decisions between each trial; the sender was not explicitly told whether the receiver called BS, and the receiver was not told whether the sender lied or not. Instead, participants only saw cumulative scores every fifth trial.

So that participants understood the payoff structure, participants completed four initial practice trials with feedback. This feedback included players' decisions, points earned, and cumulative score.

A.2.3 Analyses

Sender's Lying

Quantitative Predictions. Fitting the models—Equal Intrinsic Aversion, Unequal Intrinsic Aversion, 0^{th} Order ToM, and Recursive ToM—to the full set of data we have (i.e., the distribution over 11 messages in each of 66 circumstances for the speaker) allows quantitative model comparisons. However, such model evaluations necessarily combine a measure of the extent to which people exhibit the diagnostic patterns of behavior for each model (illustrated in Figure A.1), as well as a measure of how well the model can accommodate the non-diagnostic variability of human behavior. In typical experimental settings, subjects are not given as much freedom of choice in their behavior, and thus the behaviors are restricted to fall on fairly constrained

axes that are diagnostic of different models. In our case, we have instead opted to collect fairly unrestricted behavior, and then calculate measures that that are diagnostic of the key features of the models.

Via Linear Model. To evaluate the senders' lying behavior, we decomposed the sender's behavior into (a) their propensity to lie in each state of the world, and (b) what lies they tell when they choose to lie.

The method we used to evaluate (b) what lies people said when they chose to lie was to filter out instances when people told the truth. From there, a mixed-effect linear model was fit to the remaining false reports k^*, using the true drawn k as a predictor. The predictor k was centered at 5. Subject was included as a random intercept. This was done for each base-rate and payoff condition. From here, the intercepts of the linear fits for each condition could be meaningfully compared.

Via Mixture Model. We additionally considered (but ultimately did not include) an alternative method to evaluate the senders' behavior. This analysis attempts to control for different densities of true samples of red marbles, based on the base rate. For example, in the 80% base rate condition, there are more trials in which 8 red marbles are actually drawn, compared to in the 20% base rate condition. The evaluation assumed that (a) the propensity to lie can be formulated as a logistic function for the probability that people lied given the true number of red marbles drawn, and (b) what lies are told is a binomial function for the probability that people would report a certain value, normalizing for the probability that they would report the truth. (b) is informed by (a) and assumed to be held constant for all k in a condition.

Both analyses revealed the same qualitative pattern of behavior: that senders vary the magnitude of the lies they produced when information about the base rate changes. However, this evaluation of lying is defined as a binomial function can be held constant across all k within a given condition, which does not hold up in the Linear Model evaluation. This is because the mean and variance of lies told varies to some extent as a function of k. For example, when senders are motivated to report more red marbles, when k increases, the mean k^* also increases,

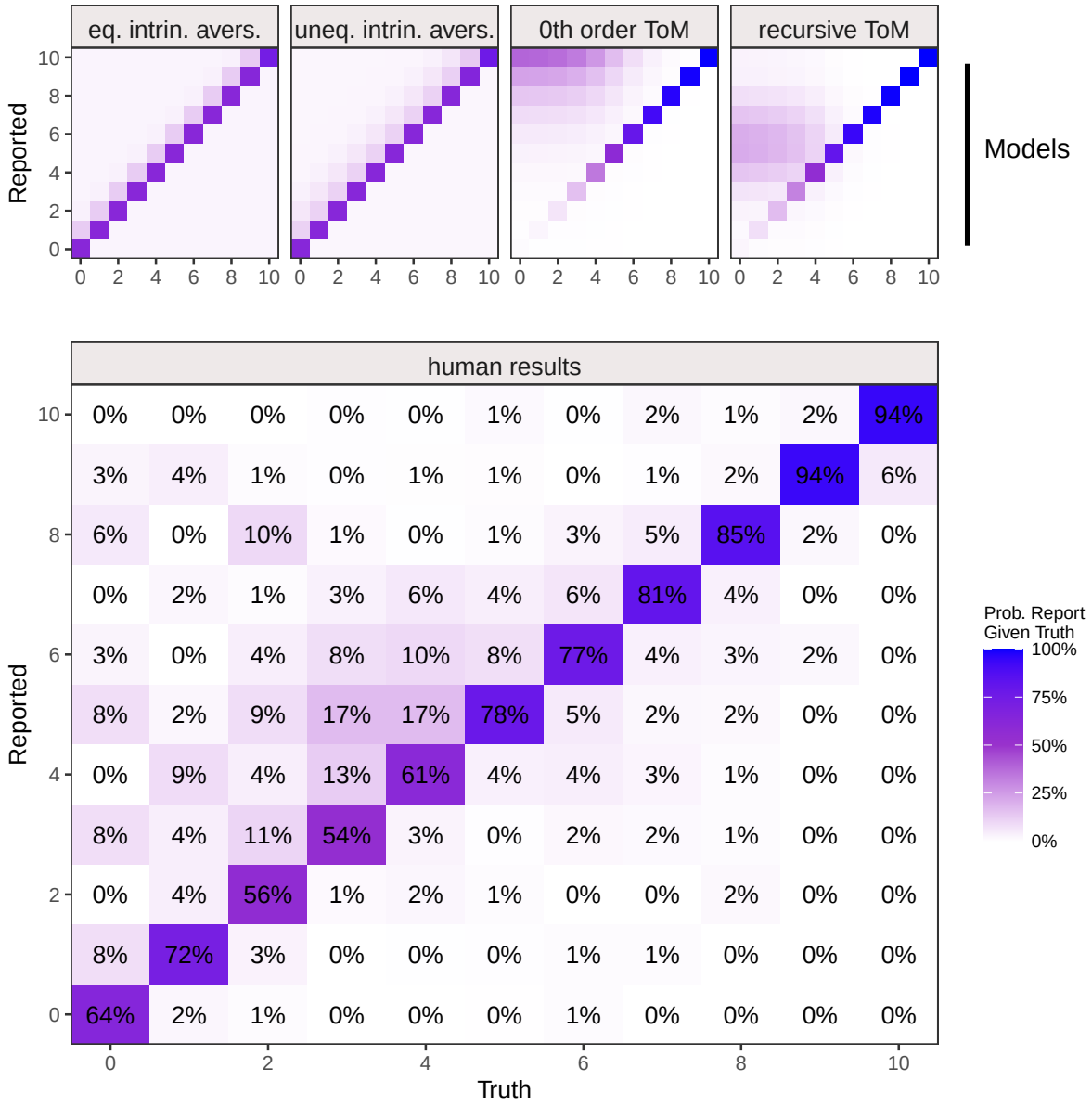


Figure A.1. Tile plots showing the model predictions (top row of plots) and human experimental results (bottom plot) of lying: the reported number of marbles sampled vs. the true number of marbles sampled. The color of the tile is the conditional probability of the report given what was true. The Recursive ToM model predicts, and we behaviorally show that people's propensity to lie depends on what was true, rather than being constant (as predicted by the Equal Intrinsic Aversion and Unequal Intrinsic Aversion models). The Recursive ToM model additionally predicts, and we behaviorally show that when people do choose to lie, they tend to hedge their lies to what is plausible from the base-rate, and not always say maximal lies, as predicted by the 0^{th} Order ToM model.

which can be judged by the positive slope of lies in the Linear Model evaluation.

Receiver's Detecting

To analyze the receivers' behavior, we fitted a vertex form quadratic logistic regression model to each condition. This analysis allowed us to capture the parabolic nature of human responses and compute shifts in BS calling rates across base rate conditions. We additionally used a lapse parameter to account for participant inattentiveness or other errors resulting in chance-level performance (Madigan & Williams, 1987). The trough parameter from the quadratic logistic fit informed the estimate of the mode of human responses' believed true reports. Meanwhile, the null account was computed from the mode of the binomial sampling distribution (10x the base rate probability). We reported the confidence intervals of the implied modes of the believed true reports for each condition.

To evaluate the fit of all models, we computed the negative log-likelihood. To show how predictions vary as a function of individual conditions, we computed the coefficient of determination r^2 for each (between subject) base rate and payoff condition. Furthermore, we examined the within subject variable—how senders vary their report as a function of the true number of red marbles drawn, and how receivers vary their BS calling behavior as a function of the opponent's report. We compared r^2 for the random opponent, recursive ToM, and recursive ToM with an aversion penalty models.

Learning

Are people learning to lie and detect lies in this task? Some task-specific learning is to be expected of participants in any new task. However, the claim that we are making is that people can spontaneously tune their lies and lie detection to others' base-rate beliefs and the payoff structure. If people can already do this spontaneously, then learning during the task should not be the primary factor driving people's behaviors.

To impede participants from tuning their behavior to the opponent via learning, our task

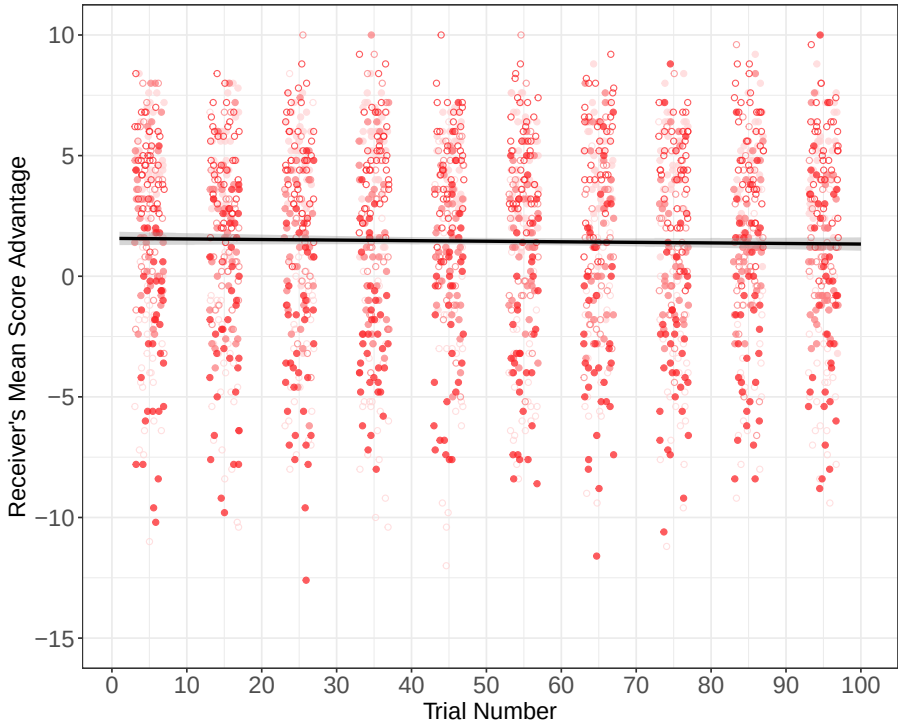


Figure A.2. Receiver's learning over trial bins. Receiver's learning is gauged as the score advantage to the player ($score_{receiver} - score_{sender}$). Base-rate conditions are shown by the color of the points, and payoff conditions are shown by whether the point is filled or not. Points represent a given participant's mean score advantage across five-trial bins. The learning curve is a linear fit to the score advantage for all trials.

provided minimal feedback to the participant (see A.2.2 Limited Feedback). Despite these efforts to minimize feedback, it may still in fact be possible that participants are tuning their behavior according to this periodic cumulative score information. Furthermore, even without attending to the scores, participants could in theory learn to tune their lies merely from observing the general distribution of their opponent's reporting behavior (without ground truth). To test that people are not learning to lie and detect lies strategically but can in fact do so spontaneously, one should expect a flat learning curve for both the sender's and the receiver's behaviors.

We use the participant's score advantage ($score_{player} - score_{opponent}$) as a proxy to gauge learning. Since participants switch between roles, they play fifty trials as each role. To examine participants' individual behaviors, we plotted their mean score advantage across ten five-trial bins. We fitted both receiver (Figure A.2) and sender (Figure A.3) learning curves as mixed effect

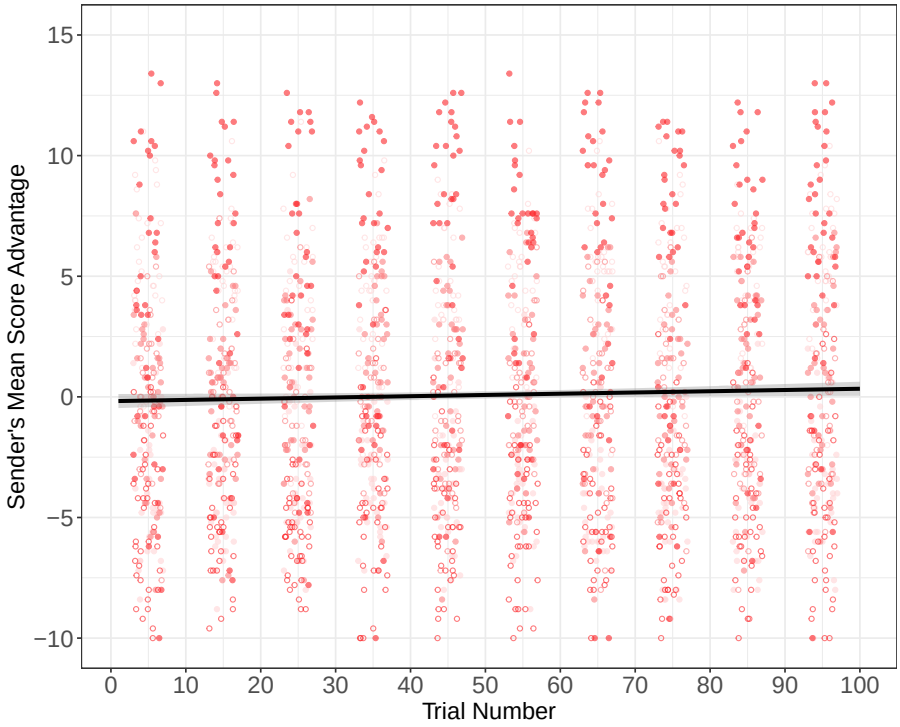


Figure A.3. Sender's learning over trial bins. Sender's learning is gauged as the score advantage to the player ($score_{sender} - score_{receiver}$). Base-rate conditions are shown by the color of the points, and payoff conditions are shown by whether the point is filled or not. Points represent a given participant's mean score advantage across five-trial bins. The learning curve is a linear fit to the score advantage for all trials.

linear regressions on the score advantage across all trials, with subject as a random intercept. For the receiver, we found no significant effect of trial number on their score advantage ($t = -1.06$, $p = 0.29$). Meanwhile, for the sender, we found a shallow but significant effect of trial number on participant's score advantage: senders increased their score advantage on average by 0.005 each trial ($t = 2.27$, $p = 0.02$). The regressions suggest that task learning may play a small but not practically significant role in how senders tune their lie, and learning plays no role in how receivers detect lies.

A.3 Experiment 2

A.3.1 Procedure

In Experiment 1, the distribution of red and blue marbles in the lying game are evenly distributed in the box and fully visible to both players. In Experiment 2, this is not always the case. The contents of the box is fully visible to the sender, but not to the receiver. The box contains a "hole" in the middle (which appears as a white rectangle), and the receiver can only see the red and blue marbles that are visible through the hole. The sender sees the white hole, so they know what is visible to the receiver, and they see the surrounding (black) box, so they know what is only visible to them (the sender). In total, there were 150 marbles in the box, offset so that marbles did not overlap too much. The inner white box contained 20 marbles and the outer black box contained 130 marbles.

We manipulated the sender's distribution of red-to-blue marbles (in the full box, white and black) and the receiver's distribution (in just the white box). Both base-rate conditions were within-subject manipulations, so all participants were equally likely to be assigned to the sender's distribution as 20%, 50%, or 80% red-to-blue marbles and the receiver's distribution as 20%, 50%, or 80% on any given trial. Marbles were sampled according to the sender's distribution. For example, if the conditions were receiver 80% and sender 20%, the true sample should generally reflect the sender 20% distribution (except in the case of one of the predetermined k trials, see A.2.2). Furthermore, the conditions when sender and receiver distributions matched were replications of base-rate conditions in Experiment 1. Since there were 9 conditions, 2 roles (sender or receiver), and a total of 100 trials, on average participants saw 5.6 trials of each condition as a given role. Additionally, senders always receive points for red, and receivers always receive points for blue.

A.3.2 Analysis

All of the materials, models, data, and analyses can be found at `https://github.com/la-oey/RationalLying/`. The raw data is available on OSF at `https://osf.io/x6rhs/`. Additionally, a visual interface for interacting with the models can be found at `https://la-oey.shinyapps.io/lierecursetomapp/`.

References

Arrow, K. J. (1965). *Aspects of the theory of risk-bearing*. Yrjö Jahnssonin Säätiö.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, *119*(3), 861–898.

Capraro, V., Schulz, J., & Rand, D. G. (2019). Time pressure and honesty in a deception game. *Journal of Behavioral and Experimental Economics*, *79*, 93–99.

Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior and Organization*, *93*, 293–300.

Goldstone, R. L., & Chin, C. (1993). Dishonesty in self-report of copies made: Moral relativity and the copy machine. *Basic and Applied Social Psychology*, *14*(1), 19–32.

Madigan, R., & Williams, D. (1987). Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations. *Perception & Psychophysics*, *42*(3), 240–249.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, *32*(1-2), 122–136.

Serota, K. B., Levine, T. R., & Boster, F. J. (2010). The prevalence of lying in America: Three studies of self-reported lies. *Human Communication Research*, *36*(1), 2–25.

van't Veer, A. E., Stel, M., & van Beest, I. (2014). Limited capacity to lie: Cognitive load interferes with being dishonest. *Judgment and Decision Making*, *9*(3), 199–206.

Verschuere, B., Köbis, N. C., Bereby-Meyer, Y., Rand, D., & Shalvi, S. (2018). Taxing the brain to uncover lying? meta-analyzing the effort of imposing cognitive load on the reaction-time costs of lying. *Journal of Applied Research in Memory and Cognition*, *7*, 462–469.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, *10*(4), 141–142.

Appendix B

Supplementary Materials to Accurate approximations about the truth from literally false messages

B.1 Human Study

B.1.1 Individual differences, or how human judge correction relates to sender bias

What mechanisms might be driving how people decide to interpret lies? There has been a recent push in the literature toward using recursive pragmatic frameworks to understand communication when speakers and listeners have misaligned goals (e.g., Oey et al., 2023; Ransom et al., 2019). A speaker reasons about how a listener might interpret an utterance influenced by the listener's beliefs and goals, and the listener in turns reasons about how a speaker ought to produce an utterance under *the speaker's* beliefs and goals. Unfortunately, people are not telepathic, so they need to conjure a model of how their opponent ought to think and behave, and decide their own behavior under that normative model.

A naïve rational heuristic might be to anchor the model of their opponent to the easily accessible model of oneself. In the marble-flipping task, a participant might just assume that the computer sender biases their lies to the same extent that she does when she is the sender, and so as the judge, she should bias-correct to the same magnitude. In other words, pragmatic frameworks suggest a systematic relationship of individual differences binding people's sender and judgment behavior (e.g., Barnett et al., 2022). Here, we evaluate the relationship between human sender biasing and human judge bias-correcting, both in aggregate and in individuals.

In aggregate

If people assume their opponent behaves like themselves, we would expect a similar (but mirrored) distribution between how human judges correct for bias and how senders bias their reports. This is what we find – the patterns of the sender's biasing and the judge's bias correcting appear to be rotationally symmetrical. Visually this highlights a similar skewed shape and spread in the bias data.

We can also compare the bias means. Figure B.1 plots the means for each condition

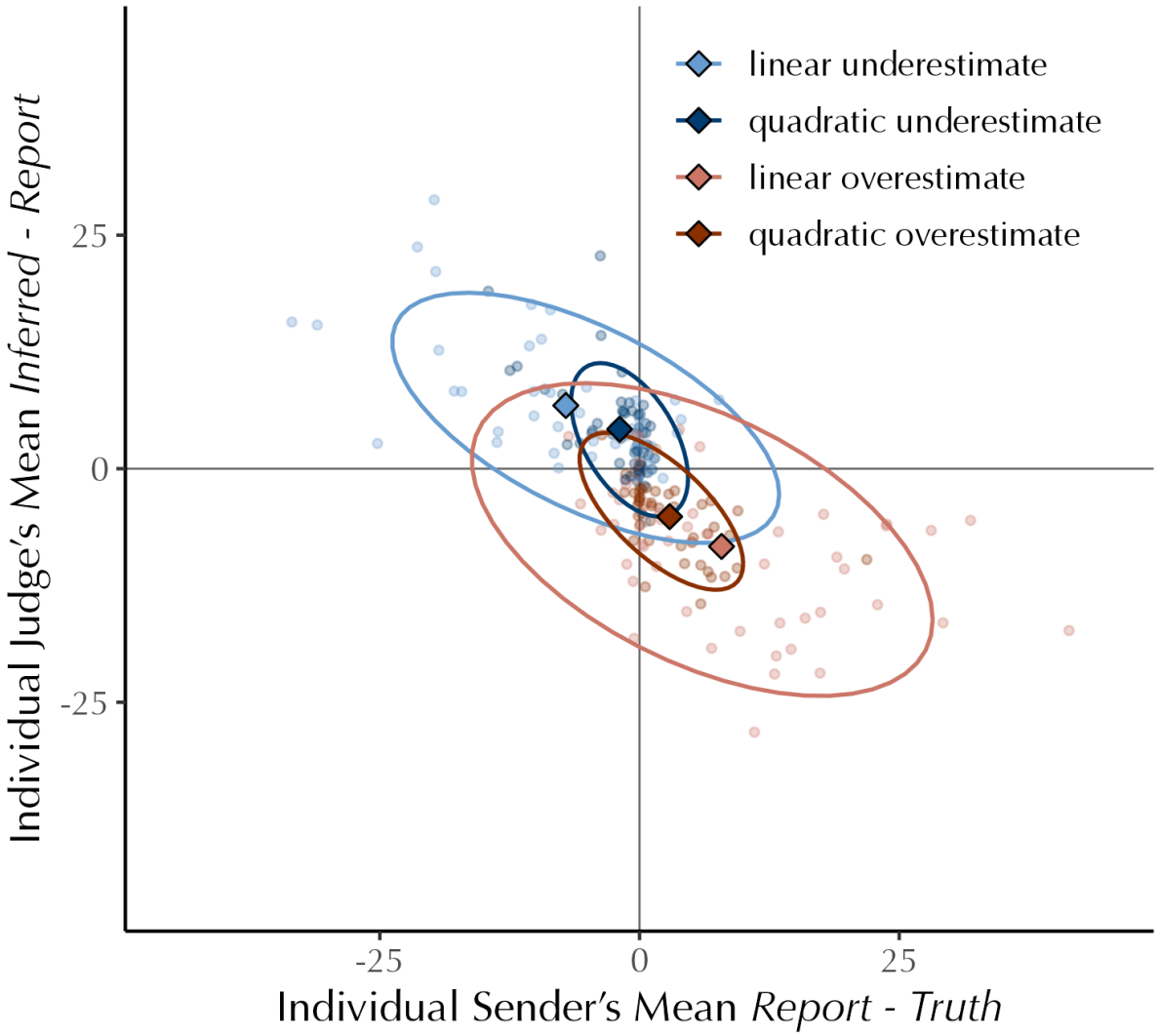


Figure B.1. Individual differences in sender and judge behavior. Sender behavior (x-axis) is summarized as the mean introduced bias ($Report - Truth$), and receiver (y-axis) is the mean bias correction ($Inferred - Report$). Scatter points represent individual participants, with rhombuses showing the mean for each condition, and ellipses showing spread. Individuals' judge behavior negatively correlates with their sender behavior, so people who produce larger lies assume their opponent also produces larger lies. The means of the *Underestimate* (and *Overestimate*) goal conditions are in the top left (and bottom right) quadrant, showing that people both biased and bias-corrected in the predicted direction. There is less spread and means are shifted toward the origin in `Quadratic` (relative to `Linear`) cost conditions.

(as rhombuses) in a 2D space, with the x-axis as senders' bias and the y-axis as judges' bias correction. If judges flip the sign of their bias, then we would expect that the *Overestimate* bias means would be located in the top left quadrant and *Underestimate* means would be in

the bottom right quadrant. Additionally, we would expect the `Quadratic` means to be shifted toward the origin, relative to the `Linear` means. These qualitative patterns are what we find. Quantitatively, examining the relationship between human senders' bias and human judges' bias by focusing on the condition means, we find a strong negative correlation of $r = -0.97$ ($t(2) = -5.46$, $p = 0.032$). Between conditions, when people produce larger bias in their reports, they correct more in the opposite direction as a judge.

In individuals

Focusing on individual differences would provide stronger evidence that people anchor their truth judgments to their own report behavior as a sender. We examined the relationship between individual senders' bias and judges' bias correction (scatter points in Figure B.1). Correcting for aggregated means, we found a significant negative correlation of $r = -0.53$ ($t(202) = -8.84$, $p < 0.0001$). An individual sender's mean bias determined 32% of the variation in their mean bias correction as a judge. These results suggest that a large contributor to what people are doing is modeling their opponent based on themselves.

B.2 Simulation

B.2.1 Implementing the probabilistic model

The senders' and judges' behaviors are outputted as a 51×51 matrices of conditional probability values $P(k_2 = K_2 | k_1 = K_1)$, where rows represent $K_1 \in \{0, 0.02, ..., 1\}$ and columns represent $K_2 \in \{0, 0.02, ..., 1\}$. Columns sum to 1. We examined senders' messages $P_S(k_{say}|k)$, judges' correction in their estimates of the truth $P_J(k_{est}|k_{say})$, and judge's truth inferences $P_J(k_{est}|k)$. Our models of lying and lie inference assume the probability of an action follows a Luce choice rule based on the expected utility of the action relative to alternative actions (Luce, 1959), with softmax parameter α set at 40:

$$P(\theta) = \underset{\theta}{softmax}(\mathrm{EV}[\theta]) = \frac{\exp(\alpha \mathrm{EV}[\theta])}{\sum_{\theta'} \exp(\alpha \mathrm{EV}[\theta'])} \tag{B.1}$$

At sufficiently large ratios of intended bias to (low) costs to say larger lies, senders are unimpeded from saying extreme messages. In these cases, the escalation goes unchecked, resulting in messages decoupled from reality. Senders always say the maximal lie, and judges always guess the minimal estimate. The measured bias for both agents plateaus and does not increase with more extreme ratios of lower costs.

B.2.2 Evaluating truth inferences

To evaluate judges' accuracy and precision of truth inferences relative to the actual truth, we first computed $P_J(k_{est}|k)$. Up until now, we characterized senders' behavior as how they generate messages based on the truth $P_S(k_{say}|k)$ and receivers' behavior as how they infer the truth from messages $P_J(k_{est}|k_{say})$. Combining the conditional probabilities with the prior probability of k and then marginalizing over all k_{say}, we jointly infer all world states of k_{est} and k. Dividing by the same, marginalized over all k_{est}, we get the conditional probability $P_J(k_{est}|k)$.

$$P_J(k_{est}|k) = \frac{\sum_{k_{say}} P_J(k_{est}|k_{say}) P_S(k_{say}|k) P(k)}{\sum_{k_{est}} \sum_{k_{say}} P_J(k_{est}|k_{say}) P_S(k_{say}|k) P(k)} \tag{B.2}$$

The prior probability of k is assumed to be uniformly distributed.

B.2.3 Measuring bias and R^2

We first measured the bias of $P_S(k_{say}|k)$, $P_J(k_{est}|k_{say})$, and $P_J(k_{est}|k)$. Bias is computed over the matrix as the conditional probability-weighted mean of errors, with the sign preserved $(K_2 - K_1)$.

We next measured the precision of truth inferences by computing the coefficient of determination R^2 of k_{est} conditioned on k. We computed R^2 via simulation. The conditional probability matrices were converted to vectors of matched pairs of k and k_{est}. This process was analogous to calculating for N instances of k, what is the average number of instances that the judge would infer the truth as k_{est} $(X = Np)$. The R^2 relationship between k_{est} and k was then

computed over the vectors. With large enough N (here, $N = 10,000$), the simulated R^2 converges to the theoretically true R^2.

References

Barnett, S. A., Griffiths, T. L., & Hawkins, R. D. (2022). A pragmatic account of the weak evidence effect. *Open Mind: Discoveries in Cognitive Science*, *6*, 169–182.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley.

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, *152*(2), 346–362.

Ransom, K., Voorspoels, W., Navarro, D. J., & Perfors, A. (2019). Where the truth lies: How sampling implications drive deception without lying. *PsyArXiv*.