Author/s:
Guo, J;Recalde, MP

Title:
Overriding in Teams: The Role of Beliefs, Social Image, and Gender

Date:
2023-04-01

Citation:
Guo, J. & Recalde, M. P. (2023). Overriding in Teams: The Role of Beliefs, Social Image, and Gender. Management Science, 69 (4), pp.2239-2262. https://doi.org/10.1287/mnsc.2022.4434.

Persistent Link:
https://hdl.handle.net/11343/333634

Overriding in teams: The role of beliefs, social image, and gender*

Joyce Guo
The University of Melbourne

María P. Recalde
The University of Melbourne

December 2021

Abstract.- To shed light on the factors that affect who speaks up in teams in the workplace, we study willingness to speak up after someone has raised an opinion. We call voicing disagreement *overriding* and study this behavior in a laboratory experiment where participants answer multiple choice questions in pairs. In a *control* treatment, participants interact anonymously. In a *photo* treatment, both participants see the photo of the person they are matched with at the beginning of the group task. Using a series of incentivized tasks, we elicit beliefs about the likelihood that each possible answer option to a question is correct. This allows us to measure disagreement and to tease apart the role of disagreement versus preferences in the decision to override ideas in teams. Results show that anonymity increases overriding. This treatment effect is driven by social image costs. Analysis of heterogeneity in behavior by gender reveals no differences between the likelihood that men and women override. However, we find some evidence that men and women are treated differently; when participants disagree with their partner, they are more likely to override a woman than a man. Preferences seem to in part explain the differential treatment of men and women. Studying group performance, we find that overriding helps groups on average, while the gender composition of teams does not affect team performance.

Keywords: lab experiment, beliefs, updating, anonymity, discrimination, promotions, gender
JEL codes: C92, D83, D91, J16, J44, J71, M12, M51

* Guo: 2019 Kinsman Student, Department of Economics (E-mail: joyceopalguo@gmail.com). Recalde (corresponding author): Department of Economics (E-mail: maria.recalde@unimelb.edu.au). This research was funded by ARC Discovery Early Career Research Grant DE190100585 and by The University of Melbourne. Ethical approval to conduct this study was provided by The University of Melbourne. We thank Boon Han Koh, Muntasha Kahn, Olga Rud, and Edwin Chan for assistance with laboratory aspects of data collection. We also thank the Department Editor, Associate Editor, and three anonymous reviewers, as well as Aaron Kamm, Joshua Miller, Nisvan Erkal, Siqi Pan, Tom Wilkening, and seminar participants at the University of Melbourne, Gender, Norms, and Economics Workshop, ESA North America, BEEC, and SITE (Experimental Economics Segment) for helpful comments.

1. Introduction

Recognizing and promoting talent in organizations is difficult. Individuals with the same ability and initial qualifications may advance at different rates because there is variation in who self-promotes, claims credit for teamwork, performs tasks that can lead to promotion, is willing to lead, and is rewarded for their effort (e.g., Babcock and Laschever 2003, Exley and Kessler Forthcoming, Isaksson 2018, Niederle and Vesterlund 2007, Babcock et al. 2017, Born et al. 2020, and Sarsons et al. 2021).[1] Another behavior that affects the recognition of talent relates to who speaks up in meetings, contributes ideas to a group, and is therefore visible to peers and managers. Empirical research and observational field data shows that not everyone speaks up at the same rate in the workplace. For example, women speak up less than men in professional environments where women are underrepresented. This includes participation in work meetings and in the classroom in MBA programs (Bursztyn et al. 2017), in academic seminars (Carter et al. 2018), in executive tech meetings (Snyder 2014), and even at the highest levels of power (Jacobi and Schweers 2017).

Several factors may affect who speaks up in teams in the workplace. Differences in preferences to speak up may exist. There may also be variation in norms of behavior which make it socially acceptable for some individuals to speak up and not others.[2] Qualifications, seniority, and experience may also play a role. They may affect not who only who speaks up in meetings, but also who is heard in a group. For example, managers and individuals with higher ranking in organizations will likely have more say and be more likely to be heard within teams. Shutting down many of these channels, laboratory experiments have studied individual willingness to speak up in horizontal environments which guarantee that everyone's voice is heard. Coffman (2014) studies the contribution of ideas in a one-shot simultaneous-move environment and shows that the stereotype associated with the task affects

[1] Since much of this variation in correlated with gender, it has been argued that this may help explain why women advance at lower rates than men in organizations. For a review of the literature on gender differences in preferences and psychological attributes and how they may contribute to the gender gap in advancement see Bertrand (2011, 2018) and Blau and Kahn (2017). An overview of the experimental literature on gender differences in preferences is provided by Croson and Gneezy (2009) and Niederle (2016). A recent review of the literature on gender differences in negotiation is provided by Recalde and Vesterlund (2020).

[2] Within the context of gender, for example, Bursztyn et al (2017) show that single women sometimes avoid career enhancing behavior that is public because the traits that are rewarded for women in the marriage market are penalized in the labor market.

individual willingness to speak up in teams. Women are less willing than men to speak up in stereotypically male environments, but more willing to do so in stereotypically female environments. [3]

To shed light on the factors that may affect who speaks up in teams in the workplace, this paper studies willingness to speak up *after* someone has raised an opinion. We call speaking up and voicing disagreement *overriding* and focus on this decision because it captures two important features of group deliberations. First, who do we listen to? In other words, how do we update beliefs in response to the ideas provided by others? Second, how do we respond to those ideas given the way we update beliefs? Teasing apart these two steps is challenging using observational field data because there are several endogenous factors at play. For example, there is endogeneity in group formation which affects how teams are formed in the workplace as well as who speaks with whom within teams. There is also endogeneity in who speaks up when and who is heard in groups. Even how long team interactions last and how overriding can occur is endogenous. While some people may be willing to publicly and directly voice disagreement, others may hesitate to do so publicly but may be comfortable doing so in private.

To overcome the endogeneity problems described above and identify the channels underlying the decision to voice disagreement in teams, we conduct a laboratory experiment. Participants answer multiple-choice general knowledge questions in pairs. We randomly vary within pairs who submits an answer first to each question. The second mover sees the choice made by their partner and decides whether to resubmit another answer on behalf of the group. We call this resubmission of a different answer overriding. Both group members earn the same sum of money if the group answer is correct. We give the second mover full authority over the group decision and through a series of tasks elicit incentivized beliefs about the likelihood that each possible answer option to a question is correct. This belief data allows us to identify the channels underlying the decision to override ideas.

There are two primary channels underlying the decision to override ideas in this experiment. The first is *beliefs;* participants will see the choice made by their partner, update their beliefs about the likelihood that each possible answer option is correct, and choose to override their partner's choice when they believe that another choice maximizes expected earnings. Beliefs therefore capture

[3] See also Bordalo et al (2019) and Chen and Houser (2019).

disagreement. The second channels is *preferences*; participants may experience a joy or disutility from overriding which causes them to behave in ways that need not maximize expected earnings. Since the second channel is likely stronger when interactions occur in public and social image concerns play a role, we conduct two treatments that vary whether interactions occur anonymously. In a *control treatment*, participants make choices anonymously without knowing who they are paired with. In a *photo treatment*, participants see the photograph of their partner at the beginning of the task. These two treatments allow us to study whether anonymity affects overriding behavior and to identify the role of observability and social image in the decision to override ideas in teams. Another question that we study is whether there is heterogeneity in behavior by gender. The photo treatment allows us to investigate whether men and women behave differently when their identity is revealed and whether men and women are treated differently.

Results show that the removal of anonymity decreases the likelihood that participants override ideas in the experiment. Analysis of the mechanisms at play reveals that beliefs (i.e. disagreement) matter and fully explain overriding behavior when interactions occur anonymously. However, beliefs cannot fully account for behavior in the photo treatment. Social image concerns play a role and cause participants to silently disagree with their partner a significant proportion of the time in the photo treatment. Analysis of heterogeneity in behavior by gender reveals two findings. First, we do not find differences in overriding behavior by gender of the decision maker. Second, we provide some evidence that the gender of the partner plays a role. When participants disagree with their partner, they are more likely to override a woman than a man. This gender difference is not driven by observable characteristics that are correlated with gender. Analysis of why women are overridden more than men suggests that the differential treatment of men and women is in part explained by preferences. Studying group performance, we find that overriding helps group on average. Teams are more likely to answer a question correctly when the second mover overrides than when the second mover does not override. We also study whether the gender composition of groups affects team performance and find no evidence that it does.

This paper contributes to the literature along several dimensions. First, to shed light on the factors that affect participation in group deliberations, we study overriding, which captures willingness

to speak up after someone has already raised an opinion. Previous work has studied gender differences in who speaks up first in groups (Coffman 2014, Bordalo 2019, Chen and Houser 2019) and differences in participation rates in environments where who speaks up when is endogenous (e.g., Born et al. 2020, Coffman et al. 2021). Prior work has not investigated group deliberations in an environment where the channels underlying the decision to speak up and override ideas can be identified.

Most closely related to our work, Isaksson (2018) conducts an experiment where participants interact in pairs and solve a puzzle taking turns to make moves within a given time period. Moves can be either good or bad, depending on whether they decrease or increase the number of steps left to solve the puzzle. The study documents a gender gap in willingness to claim credit and to revert bad moves. Men claim more credit than women and are more likely to revert bad moves. We complement this work by studying overriding behavior in an environment where the second mover has full authority over the group decision and where the length of the group interaction is fixed. This allows us to identify the channels underlying the decision to revert moves, something that to our knowledge has not been previously done in the literature. A participant who chooses not to override in Isaksson (2018) may increase the length of time participants interact with each other but need not increase the number of moves left or decrease the likelihood that the team successfully solves the puzzle.[4] The team interaction studied by Isaksson (2018) is therefore similar to a team deliberation with turn taking, where the decision to override depends not only on beliefs about the likelihood that the previous move was optimal and on overriding preferences, but also on beliefs about how the other person will respond to being overridden given the endogenous length of the team interaction.

Another contribution of our work is that in analyzing the determinants of overriding behavior, we investigate the role of gender in belief updating. Previous work has shown that there are gender differences in beliefs about own ability (Niederle and Vesterlund 2007, Coffman 2014, Bordalo et al. 2019) and in how men and women update beliefs (Mobius et al. Forthcoming, Ertac 2011, Buser et al. 2018, Coutts 2019, Coffman et al. 2019). Empirical work has also shown that gender affects how individuals form beliefs about the ability of others (Bohren et al. 2019, Sarsons 2019). We contribute

[4] Moves are not uniquely good or bad in the puzzle.

to this literature by studying belief updating within the context of team deliberations. In contrast to previous work, we find few differences in how men and women update their beliefs in this context.

More broadly, by varying whether interactions occur anonymously, our study also relates to a literature studying the role of observability and social image in individual and group choices. Laboratory experiments have shown that anonymity affects decisions in a wide array of environments such as dictator games (Hoffman et al 1996, Bohnet and Frey 1999, Andreoni and Bernheim 2009), public good games (Andreoni and Petrie 2004), trust games (Eckel and Petrie 2011), and charitable giving (Jones and Linardi 2014).[5] In the field, social image has been shown to affect behavior in domains like voting (Funk 2010, DellaVigna et al 2016), education (Burstzyn and Jensen 2015, Bursztyn et al 2016), workplace productivity (Mas and Moretti 2009) and job search decisions (Burstzyn et al 2017).[6] Our study contributes to this literature by studying the role of anonymity and social image within the context of group deliberations and overriding. We identify a cost that participants experience when they override ideas in public and interpret this cost is being driven by social image.

This paper is structured as follows. Section 2 describes the experimental design. Section 3 provides a theoretical framework we use to derive the hypotheses we test. Section 4 describes the study sample. Section 5 presents the results, and section 6 concludes.

2. Experimental design

To study willingness to speak up *after* someone has raised an opinion, we conduct a laboratory experiment where subjects are given the opportunity to resubmit an answer on behalf of the group after observing the choice made by a team member. We call this resubmission of another answer *overriding*, and study this behavior using a between-subject design with two treatments that vary whether interactions occur anonymously. Both treatments use the same session structure, which consists of three main incentivized tasks.

[5] There is also work showing no effects of anonymity alone on choices. For example, He et al (2017) show that the removal of anonymity alone does not affect cooperation rates; however, identification coupled with communication impacts behavior.

[6] See Bursztyn and Jensen (2017) for a review of this literature. Social image coupled with requests that induce social pressure have also been shown to affect decisions within the context of charitable giving (e.g., DellaVigna et al 2012, Andreoni et al 2017) and negotiation (Gago 2018).

2.1 Session structure

Task 1 consists of 12 rounds and has participants make individual choices. Tasks 2 and 3 each consists of 6 rounds and have participants interact in groups. Task 1 elicits beliefs about the likelihood that each possible answer to a question is correct. This allows us to measure individual ability and to identify cases of disagreement in subsequent tasks. Task 2 elicits individual willingness to override group decisions. Task 3 elicits beliefs about the likelihood that each answer option to a question is correct, after participants see the choice made by their partner. Task 3 allows us to study the extent to which belief updating explains overriding behavior. Either Task 1 or Tasks 2 and 3 combined were randomly selected at the end of the experiment to determine earnings. All rounds of the randomly selected task(s) were paid.

Each task is described in detail in the next subsections. While subjects knew that the experiment consisted of three tasks, they did not know what each task entailed until the relevant instructions were distributed at the beginning of each task. Instructions were read aloud to participants in every task. Instructions and decision screens are provided in the Online Appendix.

Task 1: Prior beliefs

In Task 1, participants answer 12 multiple-choice general knowledge questions in random order.[7] Each question is presented in a different round and has three possible answers options (A, B, or C). Only one answer option is correct. Participants must indicate the likelihood that each answer option to a question is correct in every round by reporting three integers between 0 and 100 which together add up to 100.

Beliefs are incentivized using the Stochastic Becker-DeGroot-Marschak Mechanism (Karni 2009). This ensures that it is incentive compatible for participants to tell the truth irrespective of their risk preferences.[8] We implement the mechanism as follows. First, one answer option (A, B, or C) is randomly drawn to determine round earnings. All answer options are equally likely to be selected. Second, an integer N, uniformly distributed between 1 and 100, is randomly drawn. If the belief

[7] These questions are presented in Online Appendix D.
[8] Burfurd and Wilkening (2018) provide an overview of the mechanism and the various prior papers that have used it to elicit beliefs. We follow the best practices they identify to reduce the role of noise and error in the data.

reported by the subject for the selected answer option is less than or equal to N, then the participant receives \$3 if the answer option is correct and \$0 otherwise. If the reported belief is instead greater than N, then the subject receives a lottery with $N\%$ chance of paying \$3. By truthfully reporting their beliefs, participants maximize their chance of earning \$3 in every round.

After reading the instructions for Task 1 and before making any decisions, participants answered a quiz aimed at evaluating their understanding. Quiz solutions were distributed to all participants, irrespective of whether they answered the quiz questions correctly, and were read aloud before Task 1 began.[9] No feedback was provided to participants regarding earnings or correct answers during Task 1.

Task 2: Group decisions

Task 2 consists of 6 rounds and has participants interact in fixed groups of 2. In each round, the group must answer one of the multiple-choice questions presented in Task 1. If the final group answer submitted is correct, both group members earn \$3 for the round. If the final group answer submitted is incorrect, both group members earn \$0 for the round.

Who speaks up first in the group is determined randomly via roles assigned at the beginning of every round. The computer randomly assigns one group member the role of *Person 1* and the other group member the role of *Person 2*. Person 1 must select one answer option to count as the group answer. The group member randomly assigned the role of Person 2, observes the answer submitted by Person 1 and chooses whether to resubmit an answer on behalf of the group. If Person 2 chooses to resubmit an answer, Person 2 can select any answer option to count as the final group answer. If Person 2 chooses not to resubmit an answer, then the answer submitted by Person 1 counts as the final answer for the group. Person 2 therefore has complete override authority over the group decision. The structure of the group decision was common knowledge between participants. At the end of each round, we revealed to both group members the final group answer submitted and who submitted it on behalf of the group. We did not reveal at any point throughout the experiment whether the final group answer submitted was correct or not.

[9] Participants were encouraged to ask questions about the task, incentives, and quiz answers before round 1 began.

Several features of the design are critical for the identification of the channels underlying the decision to override ideas in this experiment. First, random role assignment within a round ensures that who speaks up first in a group is exogenous for any given question. Second, by giving complete override authority to the second mover, the experimental design guarantees that Person 2's voice is heard. This leaves two possible channels underlying the decision to override in the experiment. The first is *beliefs*; participants may disagree with their partner and may choose to override an answer to maximize expected earnings for both players. The second is *preferences*; participants may experience a (dis)utility from overriding others' ideas which causes them to behave in a way that need not maximize expected earnings. For example, someone who experiences a disutility from overriding may sometimes choose not to override an answer even though they disagree with the choice made by their partner. Revealing who submitted the final answer on behalf of the group guarantees that both channels are active, given that participants likely experience a disutility from overriding others that is only present when the other person finds out that they are overridden. Additionally, we chose not to provide feedback about correct answers to capture the types of incentives present in team deliberations in uncertain environments in the workplace, where individuals usually know when their idea is not heard and they are overridden but it is not immediately clear who has the best idea in the group.[10] Since the questions presented in each round are different and knowing the answer to one question does not guarantee that the participant will know the answer to another question, the design used captures features of horizontal team deliberations where sometimes one team member has a better idea or insight, and other times another team member does so. Our design allows us to study who is heard and therefore not overridden in these types of teams and why.

Task 3: Posterior beliefs

Task 3 is similar to Task 2 and consists of 6 rounds. In each round, participants interact in the same groups as in Task 2 and answer one of the 6 multiple-choice knowledge questions answered in

[10]Examples of group decisions made via team deliberations in uncertain environments where the uncertainty is not immediately resolved include: hiring decisions made by committees; strategic decisions made by board of directors; marketing and creative decisions made by teams; and even policy decisions made by organizations and by the government.

Task 1 but not in Task 2.[11] Roles as Person 1 and Person 2 are randomly assigned at the beginning of every round. Person 1 makes the same decision as in Task 2 and must choose an answer option to count as the group answer. Person 2 sees the answer submitted by Person 1 but cannot resubmit an answer on behalf of the group. Person 2 instead reports beliefs about the likelihood that each answer option is the correct answer. These beliefs are private and are not revealed to Person 1.

Round earnings are determined in a slightly different manner for Person 1 and Person 2 in Task 3. Person 1 earns \$3 if the answer submitted is correct, and \$0 otherwise. Person 2's earnings can instead be determined by the choice made by Person 1 or by Person 2's reported beliefs. The two options are equally likely to determine earnings. If Person 1's choice determines Person 2's earnings, Person 2 earns \$3 if the answer submitted by Person 1 is correct and \$0 otherwise. If Person 2's reported beliefs determine her round earnings, then Person 2's earnings are determined in the same way as in Task 1. Similar to Tasks 1 and 2, there is no feedback after every round regarding correct answers or earnings and the incentive structure is common knowledge between participants.

Other procedures

At the end of the experiment we elicited risk preferences in an incentivized way, using a multiple price list (Holt and Laury 2002). After collecting data on risk preferences, we also asked participants to answer a non-incentivized survey. This survey collected socio-demographic information such as gender, age, ethnicity, and unincentivized beliefs regarding what gender had an advantage in the main experimental tasks.[12]

2.2 Treatments

To study the role of disagreement, social image, and gender in the decision to override ideas in teams, our experiment uses a between-subject design with two treatments that vary whether participants interact anonymously. In a *control treatment*, participants interact anonymously in the manner described

[11] Questions were presented to participants in random order in Tasks 1-3. While the draw was independent for each participant in a session in Task 1, there were two sequences assigned to groups in Tasks 2 and 3. These sequences were created from a random draw of questions and differed only in that they flipped the questions assigned to Tasks 2 and 3. The exact order of questions presented to subjects is provided in Appendix Table 1.

[12] We also elicited unincentivized beliefs regarding whether Native English speakers and Australian born participants had an advantage in the task. In one of the treatments described in section 2.2, we additionally asked participants whether they knew, had communicated, or were friends their partner prior to the experiment.

above. In a *photo treatment*, we take participant photos at the beginning of the session and show subjects the photo of the person they are matched with at the beginning of Tasks 2 and 3. All participants in the photo treatment therefore know the gender and identity of the person they are matched with. They are also aware that their partner knows who they are.[13]

Several factors change between the two treatments which may affect individual willingness to override ideas in the experiment. For example, photographs reveal individual characteristics that may affect the way individuals update beliefs about the likelihood that the answer chosen by their partner is correct. By making choices observable within teams, the photo treatment also activates overriding preferences such as those driven by social image. Using Task-3 belief data, we can tease apart how much of the overall treatment effect is driven beliefs versus preferences. We describe this in the next subsection. Photographs may also increase the sense of scrutiny that participants perceive from the experimenter. Using Task-1 data we can investigate whether there is an effect of photographs alone on choices.

Within the context of gender, the photo treatment reveals own and partner characteristics and in doing so activates gender stereotypes as well as gender-specific preferences that may not be present when interactions are anonymous. Stereotypes may affect the way individual update beliefs and thus the likelihood that participants ultimately agree or disagree with their partner. Gender-specific preferences, on the other hand, refer to differences in the joy or disutility participants experience when they override a man or a woman. Using Task-3 beliefs, our design can separate how much of the overriding of men and women is driven by beliefs versus preferences. We cannot distinguish the extent to which stereotypes affect belief updating or how much of the overriding preferences are driven by a specific type of preference. We discuss this in the next subsection. Since gender is not the only

[13] We do not conduct an additional treatment where gender is revealed and anonymity is maintained for several reasons. First, related papers document gender differences in environments where participants either see the photo of the person they are matched with or interact with each other face to face (e.g. Isaksson 2018, Born et al 2020). Second, using photos minimizes concerns about experimenter demand effects since several traits in addition to gender are revealed via photographs. Experimental studies that reveal gender in anonymous decision environments include Bohnet et al (2016), Babcock et al (2017), and Bordalo et al (2019), Charness et al (2020), Manian and Seth (2021).

observable characteristic revealed via participant pictures in the photo treatment, we also examine the extent to which other observable characteristics play a role in the results section.

3. Theoretical framework and hypotheses tested

Let $u_i(\cdot)$ represent the utility person i derives from the group decision, which can be x_i if i overrides j, and x_j otherwise. Person i's overriding decision given j's choice in question q is determined by the expected utility associated with each possible answer option. This expected utility is given by

$$E\left[u_i(x_i, x_j)\right] = \begin{cases} \$3p_i\left(x_i \middle| x_j\right) - c_{ij,T} & if\ override \\ \$3p_i\left(x_j \middle| x_j\right) & otherwise \end{cases}$$

$p_i\ (x_k \mid x_j\ )$ denotes i's belief that k is the correct answer given j's choice. $c_{ij,T}$ is a constant that captures the disutility or joy i experiences when overriding, which may vary with the identity of i and j and with treatment $T = \{Control, Photo\}$.[14] Task 2 identifies whether $\$3p_i\left(x_i \middle| x_j\right) - c_{ij,T} \geq \$3p_i\left(x_j \middle| x_j\right)$ and thus what choice is optimal given i's beliefs and preferences. Task 3 elicits $p_i\left(x_i \middle| x_j\right)$ and $p_i\left(x_j \middle| x_j\right)$, which allows us to construct a counterfactual of the rate of overriding we would observe if $c_{ij,T} = 0$. We use this counterfactual to empirically identify $c_{ij,T}$.[15]

Since the two treatments vary the anonymity of participant choices, they vary the social image incentives at play. We hypothesize that participants experience a disutility from overriding on average and that social image concerns play a role, making the disutility higher when interactions are not anonymous (i.e. $0 \leq c_{ij,Control} < c_{ij,Photo}$). This implies that controlling for beliefs, overriding rates are higher when participants interact anonymously.

- *Hypothesis 1:* Overriding rates are *higher* in the control treatment than in the photo treatment.

[14] $c_{ij,T}$ could be endogenized to capture reciprocity concerns present in the repeated game or dynamic features of decision making. Because teams answer a different question in each round and being overridden in the past does not significantly affect overriding behavior in our data, we do not write a more complex dynamic model of overriding.

[15] To empirically identify $c_{ij,T}$, we assume that belief updating is constant across Tasks 2 and 3 within treatment. If participants are more likely to give their partner the benefit of the doubt in earlier rounds than in later rounds, then the counterfactual $c_{ij,T}$ that we estimate within treatment could be biased upward. Since this bias, if present, is constant across treatments, comparison of $c_{ij,T}$ estimates across treatments will inform us of the size of this bias. We chose not to vary the order in which Tasks 2 and 3 were conducted because eliciting updated beliefs before overriding decisions could prime behavior in the overriding task and make it more dependent on beliefs than on preferences.

A corollary of hypothesis 1 is that beliefs alone cannot explain overriding behavior, at least in the photo treatment where social image concerns play a role.

- *Hypothesis 2:* The overriding rate predicted by posterior beliefs when $c_{ij,T} = 0$ is *higher* than the overriding rate observed within treatment. [16]

Moving onto gender differences, based on prior work on team tasks (e.g. Coffman 2014, Isaksson 2018, Born et al. 2020), we hypothesized that overriding rates might differ between men and women. Specifically, we thought men might override more than women when participants know who they are interacting with.[17] This may be due to differences in beliefs or differences in the (dis)utility men and women experience when they override others.

- *Hypothesis 3:* Men override *more* than women in the photo treatment.

The experiment also allows us to examine whether men and women are treated differently. Based on the literature on interruptions (e.g. Smith-Lovin and Brody 1989, Jacobi and Schweers 2017), discrimination (e.g., Bohren et al. 2019, Sarsons 2019), academic seminar dynamics (Dupas et al 2021), and influence in group deliberations (Born et al. 2020), we hypothesized that women may be overridden more than men. This can be due to the effect that j's gender has on i's beliefs or due to variation in the (dis)utility i derives from overriding men versus women.

- *Hypothesis 4:* Men are overridden *less* than women in the photo treatment.

We do not make predictions regarding the interaction of own and partner gender in the photo treatment and therefore consider such analysis exploratory in the paper.

4. Study sample

The experiment was programmed in Z-Tree (Fischbacher, 2007) and conducted at the University of Melbourne Experimental Economics Laboratory. Participants were recruited from the subject pool of university students using recruiting software ORSEE (Greiner, 2015). Only participants

[16] If participants experience a joy from overriding on average and social image incentives play a role (i.e. $c_{ij,Photo} < c_{ij,Control} \leq 0$), then the theoretical predictions in hypotheses 1 and 2 flip.
[17] A pilot study conducted to select gender neutral questions for the study had participants interact anonymously and showed no gender differences in the likelihood to override group decisions.

30 years old or younger, who self-identify as male or female, and who had no prior experience with similar experiments were invited to participate.[18]

A total of 254 individuals participated in the study. Eleven sessions of the experiment were conducted. Session size ranged between 18 and 24 participants. A total of 94 subjects participated in the control treatment and 160 subjects participated in the photo treatment.[19] Photos in the non-anonymous treatment were taken to the count of three, in front of a blue screen, after participants gave consent to participate in the study and before instructions were distributed. We did not take participant photos in the control treatment. Treatment effects are therefore the combined effect of taking participants' photos, showing them the image of their partner, and having their partner see their image. However, as section 5 will show, we do not find evidence that taking participants photos alone affects behavior in the experiment. Sessions lasted between 1.5 and 2 hours.[20]

Although we invited an equal number of men and women to participate in all sessions, gender was slightly unbalanced in the final sample of participants; a total of 118 men (46%) and 133 women (52%) participated.[21,22] The share of female participants in a session ranged between 50-55% in the control treatment, and between 46-58% in the photo treatment. Ethnicity is mixed, 56% of participants

[18] Because data on the nature of prior experiments conducted in the lab was incomplete, we limited our recruitment to subjects who had participated in 5 or fewer experiments.

[19] Two thirds of the sample received the photo treatment because there is variation in own and partner characteristics that are observable in the photo treatment. We aimed to have an equal number of participants of each gender matched with an unknown partner in the control treatment, and with a man and a woman in the photo treatment. Our target sample size was 288 or more participants. Using the group as the unit of observation and the overriding rate observed in the pilot study (Mean: 0.2479, SD: 0.1167), we planned to have 80% power to detect a treatment effect of 5.82 percentage points or more at the 0.05 level of significance. Similarly, we aimed to be powered to detect a gender difference of 8.83 percentage points or more within the photo treatment. Unfortunately, we fell short of meeting our recruitment goal and thus discuss power limitations when appropriate in the results section.

[20] This includes consent and payment procedures. Task-1 instructions and the associated comprehension quiz took approximately 30 minutes.

[21] A socio-demographic survey conducted at the end of the experiment asked participants to indicate their gender. Even though we recruited only participants who self-identify as male or female, 3 participants did not identify as male or female in the survey. We exclude data from their groups from the main analysis of results. Our findings are robust to including this data in the analysis. In Online Appendix E, we also use a perceived gender variable coded from the photos of participants in the photo treatment. Perceived gender is equal to self-reported gender among all participants who self-identify as male or female in the photo treatment.

[22] There were 3 cases of attrition in the experiment, all men in the photo treatment. One participant declined to participate after reading the consent form and before having their photo taken. Because we overrecruit participants for each session, we were able to replace this participant with another subject recruited for that same session. Another participant was dismissed from the experiment during Task 1 because he was using his cell phone. This was prohibited throughout the duration of the session. Given that Tasks 2 and 3 involved group decisions, the subject's randomly assigned partner also had to be dismissed at the beginning of Task 2. Similar procedures were used to monitor cell phone use across all sessions and treatments.

are East or Southeast Asian, 15% are Caucasian, and 29% are from other ethnic backgrounds.[23] Appendix Table 2 reports summary statistics by gender and treatment. Because there is some imbalance in subject characteristics by gender and treatment, we report estimates with and without a full set of controls in our analysis.

5. Results

We begin our analysis of results by studying treatment effects in the pooled sample of men and women. Sections 5.1 and 5.2 present the raw data by treatment. In Section 5.2 we also formally test hypothesis 1, whether the photo treatment decreases overriding, using models that control for individual characteristics as well as question or round effects. Section 5.3 examines the channels underlying the decision to override in the experiment. Section 5.4 studies heterogeneity in behavior by gender of the decision maker and gender of the partner. Section 5.5 studies group performance, and Section 5.6 describes the robustness checks we perform.

5.1 Prior beliefs by treatment (Task 1)

To assess the difficulty of each question and generate an individual ability measure, we use Task-1 belief data to construct two measures of ability. The first measure uses Task-1 beliefs to predict the choice that each participant would have made if we had asked them to choose one answer option.[24] We construct this predicted choice by identifying the answer option that a participant assigns the highest probability mass to in any given question. We then construct a dummy variable indicating whether the predicted choice is correct.[25] Our first measure of individual ability is the proportion of predicted choices that are correct across all questions. Our second measure uses instead the raw belief assigned

[23] We use self-reported ancestry/ethnicity and photos from the non-anonymous treatment to generate ethnicity variables. Participants who report their ancestry as European, from Australia or New Zealand, and the US or Canada are classified as Caucasian. The 'other' category includes Africa, Central and South Asia, Latin America, the Middle East, and Other. Indigenous Australian or Torres Islander was a possible answer category, but no one chose it in our data. In section 5.4, we check whether using a perceived ethnicity variable coded from the photos rather than self-reported ethnicity makes a difference. While the mapping from self-reported to perceived ethnicity is not one to one, results are similar.

[24] Appendix Figure 1 shows distribution of prior beliefs by treatment, question, and answer option. While there is variation in reported beliefs across questions and answer options, the distribution of beliefs within question and answer option are very similar.

[25] We allow for ties. A predicted choice is correct if one of the answer options that received the highest probability mass was correct. By design 3-way ties were impossible, beliefs were elicited using integers between 0 and 100 and there are three answer options.

to the correct answer in each question to compute the mean belief participants assign to the correct answer across all questions.

The left panel of Figure 1 shows the proportion of predicted choices that are correct by treatment. Two things are worth highlighting. First, questions are difficult; predicted choices are correct only 38% of the time. Second, there are no differences in ability across treatments (Mean Control: 38.4%, Mean Photo: 37.7%, p=0.689). Comparing the proportion of predicted choices that are correct to the proportion that would be correct by random chance (1/3), we can reject the null hypothesis of random guessing on average ($p<0.01$ in each treatment).[26]

The right panel of Figure 1 shows the mean belief participants assign to the correct answer across all questions. Again, there are no differences in ability by treatment. To check whether participants are randomly guessing, we examine the beliefs reported by participants in each question. Panel A of Appendix Figure 2 shows the average belief assigned to the correct answer by question and treatment. We can reject the null hypotheses that beliefs are on average equal to 100/3 in 10 out of 12 questions.[27] We can also use the raw beliefs provided by participants in Task 1 to identify possible cases of three-way ties; situations in which the distribution of beliefs is 34-33-33 across answer options. Doing so reveals three-way ties in at most 9.11% of cases. The raw data therefore suggests that there is not a large proportion of random guesses even though the mean belief assigned to the correct answer across all questions is on average not different than 100/3.[28] Panel B of Appendix Figure 2 shows the distribution of individual ability scores by treatment. We restrict attention to the mean belief assigned to the correct answer; however, results are similar if we instead show the proportion of predicted choices that are correct. We find no evidence of differences in ability across treatments. This rules out the possibility that taking participant photos affects behavior in Task 1 and suggests that any treatment

[26] All p-values reported in the paper are from two-sided tests.
[27] We cannot reject the null hypothesis of random guessing on average in questions 5 and 10 only, which appear in round 3 (see Appendix Table 1). For all other questions $p<0.05$.
[28] This compression of beliefs around 100/3 is consistent with a "pull-to-center" effect that makes reported beliefs more conservative than the true underlying beliefs. We therefore use the raw belief assigned to the correct answer to construct ability scores and proxies of initial disagreement size only. Choices predicted by beliefs are used to identify the role of beliefs (disagreement) versus preferences using data from Tasks 2 and 3. See Danz et al (2020) for a discussion of belief misreporting when binary lotteries are used to incentivize truth-telling.

effect observed in Task 2 is generated by non-anonymity rather than differences in experimental procedures across treatments.

The difficulty of questions was expected given that by design we wanted to study cases of disagreement. A natural question to ask is whether the beliefs elicited in Task 1 are consistent with the choices made by participants in subsequent tasks. To study this question, we check whether the choice predicted by beliefs using Task-1 data is equal to the actual choice made by participants when assigned the role of Person 1 in Task 2. Appendix Figure 3 shows that there is a high degree of consistency in the data, which does not vary with treatment. Participants' predicted and observed choices coincide 85% of the time. This rate of consistency increases to 89% when almost ties are counted as consistent choices. We define as almost ties as cases in which there is a 1-percentage point difference between the belief assigned to the correct answer and the belief assigned to the answer option that received the highest probability mass (e.g. 34-33-33). These statistics confirm the high quality of our belief data, which we use to identify cases of disagreement in subsequent tasks and to study why participants override ideas in the experiment.[29] We describe these results in the next subsection.

5.2 Overriding by treatment (Tasks 2)

We define disagreement as any case in which the predicted choice made by Person 2 in Task 1, using prior belief data, does not coincide with the choice made by their partner in Task 2. The left panel of Figure 2 shows disagreement rates by treatment. We succeeded in designing a study where participants disagree with each other, Person 2 disagrees with Person 1 nearly 50% of the time in each treatment. Because participants should not override their partner by resubmitting a different answer when they agree with them, the disagreement rate we identify using prior belief data provides an upper bound of the rate of overriding we can expect to observe in the experiment.[30,31]

[29] Further evidence of the high quality of our belief data is provided by the results of another experiment we describe in Online Appendix E, where participants had to choose one answer option when answering Task-1 questions rather than report beliefs. We observe a similar proportion of correct choices than what is reported in the Panel A of Figure 1 (38%).

[30] Theoretically, participants may override their partner when they agree with them if they experience a joy from overriding that is strong enough to outweigh the monetary costs of resubmitting a different answer and thus not maximizing expected earnings. We find no evidence that participants experience such a joy from overriding in the data.

[31] Disagreement may vary in size. We use belief data from Task 1 to construct a disagreement size variable, which is equal to the difference between the highest belief Person 2 assigned to an answer option in Task 1 and the belief Person 2 assigned to the answer option chosen by Person 1. 0 represents cases of agreement. Appendix Figure 4

The right panel of Figure 2 shows overriding rates by treatment. Even though participants initially disagree with their partners nearly 50% of the time, they override only 25.2% of the time in the control treatment and 16.7% of the time in the photo treatment. This difference in overriding rates is statistically significant (p=0.01) and represents a 34% reduction in the probability of overriding generated by the photo treatment. To gain precision and control for question specific effects, we formally test hypothesis 1 by estimating a linear probability model of overriding that includes question dummies and restricts attention to cases where participants disagree with each other.[32] We estimate this model using ordinary least squares.

$$Override\_t2_{igq} = \beta_0 + \beta_1 Photo_{ig} + \gamma_q + X'\theta_i + \epsilon_{ig} \quad (1)$$

$Override\_t2_{igq}$ is an indicator for whether subject i overrides in Task 2. q denotes question and g denotes group. $Photo_{ig}$ is an indicator for the photo treatment, γ_q are question dummies, X is a vector of individual-level controls, and ϵ_{ig} denotes standard errors clustered at the group level. Individual-level controls include gender, age, ethnicity, field of study, and native language indicators as well as an individual ability score which is equal to the mean belief a participant assigns to the correct answer across all questions in Task 1.

Estimates of model (1) are presented in columns 1-3 of Table 1. Column 1 shows that participants are 13.7 percentage points (31%) less likely to override their partner when they initially disagree with them in them in the photo treatment relative to the control. Column 2 adds controls for individual characteristics and shows that the probability of overriding increases with individual ability, the only coefficient from the vector of controls that we report. Column 3 adds disagreement size, history of play, and risk preferences as regressors. While individual ability and disagreement size affect the likelihood that a participant overrides in the experiment, the additional control variables do not explain the treatment effect we observe. The coefficient on disagreement size indicates that going from a prior

shows the distribution of disagreement size by treatment. Disagreement ranges in size from very little disagreement (1 percentage-point difference in prior beliefs, 5% of all cases) to full disagreement (100 percentage-point difference, 14% of all cases). The mean and median disagreement size conditional on disagreeing is 53.92 and 50 respectively.

[32] Results are similar if we use the full sample of observations. We discuss this in section 5.6.

disagreement size of 1 (slightly disagreeing) to 100 (completely disagreeing) increases the probability of overriding by 48 percentage points on average.

The main model of overriding that we estimate includes question fixed effects because there is variation in difficulty across questions and because our design and implementation procedures cannot distinguish between question and round effects. Each question appears in one round only (see Appendix Table 1). Nevertheless, to study the dynamics of overriding across rounds, we also estimate a linear probability model of overriding that instead of including question dummies controls for round and question sequence. Estimates presented in columns 4-6 of Table 1 show similar results. However, the coefficient on round of play is positive and statistically significant, indicating that the probability of overriding increases by 3 to 5 percentage points on average with each round of play. This result suggests that participants are more likely to voice disagreement the more the interact with their partner. We do not find evidence that being overridden in a previous round affects the probability of overriding or that risk preferences play a role.[33] Appendix Table 3 presents estimates of Table 1, which do not restrict attention to cases of disagreement and shows similar results.[34]

Estimates of the relationship between disagreement size and overriding presented in Table 1 reveal that beliefs about the likelihood that the other person made the correct choice matter and affect the decision to override in the experiment. However, the disagreement size variable that we use is constructed using Task-1 belief data and therefore captures prior beliefs only. When making the decision to override in Task 2, participants make choices after seeing the choice made by their partner and updating their beliefs accordingly. Therefore, we use belief data from Task 3 to study the extent to which belief updating explains overriding behavior in the experiment. We describe this data in the next subsection.

5.3 Overriding observed versus predicted by beliefs (Tasks 2 and 3)

To understand how much of the reduction in overriding generated by the photo treatment is driven by disagreement versus preferences, we use belief data from Task 3 to construct a predicted

[33] This result could be task specific and may differ if another measure of risk preferences was used. See, for example, Crosetto and Filippin (2015), Filippin and Crosetto (2016), and Eckel and Grossman (2008).
[34] See also Appendix Figure 6, which plots the mean overriding rate by round and treatment.

override variable. This variable indicates whether a subject assigns the highest probability mass in Task 3 to a choice different than the one chosen by Person 1. We use this predicted choice to estimate equation 2 and investigate whether the rate of overriding observed versus predicted by beliefs differs between participants who answer the same question in Tasks 2 and 3.

$$Override\_t2\&3_{igq} = \beta_0 + \beta_1 Photo_{ig} + \beta_2 Predicted_{ig} + \beta_3 PredictedXPhoto + X'\theta_i + \gamma_q + \epsilon_{ig} \quad (2)$$

Variable $Override\_t2\&3_{igq}$ is equal to $Override\_t2_{igq}$ in Task 2 and to predicted override in Task 3. $Predicted_{ig}$ is a Task-3 dummy, all other notation is the same as in equation (1). We report estimates of β_2 and β_3 in Table 2 only. β_2 identifies the average cost (disutility) that participants experience when they override in the control treatment. β_3 identifies the additional disutility that participants experience in the photo treatment, which we interpret as being driven by social image. $\beta_2 + \beta_3$ is the total disutility participants experience when they override in the photo treatment.

Results presented in columns 1-3 of Table 2 show that the rate of overriding predicted by posterior beliefs fully explains overriding behavior in the control treatment, but not in the photo treatment. Column 1, for example, shows that there is a 7.1 percentage point gap between observed and predicted overriding rates in the control treatment, which is not statistically different from zero. The coefficient on PredictedXPhoto shows that this gap is 18.7 percentage points larger in the photo treatment. This supports hypothesis 2, which states that total overriding costs are larger in the photo treatment than in the anonymous control treatment. Results are similar once controls for individual characteristics are added to the regression (column 2), and become marginally significant once disagreement size, history of play, and risk preferences are included as regressors (column 3). Estimates of β_3 from column 3 suggest that the additional disutility participants experience when they override in the photo treatment may be in part explained by risk preferences, disagreement size, and history of play. Nevertheless, the overall pattern of results is similar across all columns. There is a large and robust gap between observed and predicted overriding rates in the photo treatment and estimates of the distaste for overriding that is driven by social image (β_3) are larger in magnitude than the total distaste for overriding identified in the control treatment (β_2). Columns 4-6 show that results are similar if we

include round and question sequence controls rather than question dummies. Appendix Table 4 presents estimates of Table 2 which include cases of agreement in the estimation sample and shows similar results.

Since we observe overriding in Task 2 and elicit beliefs in Task 3, it is possible that our estimates of the disutility participants experience from overriding underestimate the importance of belief updating within treatment and thus overestimate the role of preferences. This could occur if beliefs become less noisy as participants interact with each other across rounds, or if participants give their partner the benefit of the doubt in Task 2 but not in Task 3. Since this bias, if present, is constant across treatments, it does not affect our estimate of β_3—the distaste for overriding that is unique to the photo treatment and we interpret as being driven by social image.

5.4 Heterogeneity by gender

Sections 5.2 and 5.3 have examined treatment effects and the channels underlying the decision to override ideas in the pooled sample of men and women. In this section we study hypotheses 3 and 4, which predict differences in overriding behavior by gender of the decision maker and gender of the partner in the photo treatment. Before proceeding to study these questions, we first examine whether there are gender differences in ability. Figure 3 plots mean ability by gender using the two ability measures described in section 5.1: (1) the proportion of predicted choices that are correct, and (2) the mean belief participants assign to the correct answer across all questions. Both panels show that some gender differences in ability exist; the proportion of predicted choices that are correct (left panel) is 39.6% for men and 36.5% for women (p=0.088), while the mean belief assigned to the correct answer (right panel) is 35.4 for men and 32.7 for women (p=0.03).[35,] There is thus a slight male advantage in the task, which is consistent with reports provided by participants when they answered a survey question at the end of the experiment.[36]

[35] Appendix Table 2 shows ability scores by gender and treatment. Gender differences are not statistically significant within treatment (P=0.133 and 0.101 in the control and photo treatments respectively). Appendix Figure 7 shows distribution of Task-1 beliefs by gender of the decision maker, question, and answer option in the pooled sample of treatments.

[36] Participants were asked the following survey question to identify whether there was a perceived gender advantage in the Task: "If 100 participants were to answer the same multiple-choice knowledge questions that you answered in Tasks 1-3 today, who do you think would answer more questions correctly?" Possible answer options included: (1) Definitely men, (2) Probably men, (3) No gender difference, (4) Probably women, and (5)

Panel A also shows that the proportion of predicted choices that are correct is significantly higher than 1/3 for both men and women ($p<0.05$ for men and women). In Panel A of Appendix Figure 8 we show the mean belief assigned to the correct answer by question and gender. We can reject the null hypotheses that beliefs are on average equal to 100/3 in 10 out of 12 questions for both men and women. Using raw beliefs to identify possible cases of three-way ties (i.e. choices in which the distribution of beliefs is 34-33-33), we find that at most 11.3% of choices made by men and 7.2% of choices made by women may reflect three-way ties.[37] Panel B of Appendix Figure 8 plots the distribution of individual ability by gender and Appendix Figure 9 shows the rate of consistency between Task-1 beliefs and Task-2 choices by gender and treatment. We find no differences in the rate of consistency by gender.

Overriding by gender and treatment (Task 2)

We now proceed with the presentation of the raw overriding data, before formally testing hypotheses 3 and 4. Figure 4 shows overriding rates by gender of the decision maker and treatment. The graph does not restrict attention to cases of disagreement, and as such is directly comparable to Figure 2. It shows that the negative photo treatment effect is present in the subsample of men and women. Men override 27.8% of the time in the control treatment and 17.3% in the photo treatment ($p=0.044$), while women override 22.9% of the time in the control treatment and 16.1% of the time in the photo treatment ($p=0.085$).[38] Comparison of overriding rates by gender and within treatment reveal no differences.

Figure 5 plots overriding rates by gender of the partner interacted with treatment. We do not distinguish between men and women in the control treatment, because the gender of the partner is unknown to participants in when interactions occur anonymously.[39] Figure 5 reveals that the negative

Definitely women. Even though most men and women reported no perceived gender difference (82.9 and 88.6% respectively), more men than women indicated that there is a male advantage in the tasks (Mean Men: 2.91, Mean Women: 3.03, t-test $p=0.0195$, Fisher exact test $p=0.066$). The mean answer provided by men is statistically different from 3 ($p=0.024$), while the mean answer provided by women and by the pooled sample of participants is not ($p=0.348$ and 0.331 respectively).

[37] This difference is marginally significant. An OLS regression of a possible three-way-tie indicator on a female dummy with standard errors clustered at the subject level reveals $p=0.064$.

[38] The rate of disagreement is 55.6% and 54.7% for men and 46.5% and 52.2% for women in the control and photo treatments respectively.

[39] This is also the type of model we use to identify the role of beliefs versus preferences using Task-2 and Task-3 data in the next subsection. Appendix Figure 10 shows overriding rates by partner gender and treatment.

photo treatment effect observed in the pooled sample of men and women is present when participants are paired with both men and women. The overriding rate drops from 25.2% when participants are paired with an unknown partner in the control treatment, to 14.2% when they are paired with men in the photo treatment and to 18.9% when they are paired with women in the photo treatment (p=0.03 and 0.094 respectively). The graph also suggests that women are overridden more than men in the photo treatment; however, this difference is not statistically significant (p=0.183).

To gain precision and control for question-specific effects, we once again restrict attention to cases of disagreement and estimate linear probability models of overriding that include question dummies, cluster standard errors at the group level, and systematically add controls to the regressions to account for variables that may be imbalanced across genders. To test hypothesis 3, we first examine whether the gender of the decision maker affects overriding behavior within treatment. Results presented in Table 3 show that we cannot reject the null hypotheses of no gender differences in overriding rates within treatment. Columns 1-3 suggest that women are slightly *less* likely to override than men, while columns 4-6 suggest that they are *more* likely to do so. However, none of the estimates are statistically different from zero. Results are similar if we include cases of agreement in the estimation sample as shown in Appendix Table 5.

The lack of support for the hypothesis that men override more than women in the photo treatment is surprising for several reasons. First, given the results previously documented in the literature (e.g. Coffman 2014, Isaksson 2018, Born et al 2020), we expected gender differences to emerge in the photo treatment. The difference in results could be due to type of decision environment we study but could also be driven by the different subject pool we use. Second, there is a slight male advantage in the task which should push men to override more than women. Importantly, male participants are aware of this advantage, so gender stereotypes could play a role in the same way they do in other studies in this literature. Third, even though we are not well powered to detect small differences in the likelihood that men and women override, the results documented in column 6 of Table

3 allow us to rule out an average gender difference of more than 3.3 percentage points in the hypothesized direction—men overriding more than women.[40]

We now turn to the analysis of hypothesis 4, which asks whether men and women are overridden at different rates when their identity is known in the photo treatment. To test this hypothesis, we estimate a linear probability model of overriding that includes two interacted variables as the independent variables of interest: Photo X male partner, and Photo X female partner. The omitted category is the unknown partner in the control treatment. Results presented in columns 1 of Table 4 show that conditional on initially disagreeing with their partner, participants are 18.9 percentage points less likely to override a man in the photo treatment than an unknown partner in the control treatment. Participants are also 8.8 percentage points less likely to override woman in the photo treatment than an unknown partner in the control treatment. This last coefficient, however, is not statistically different from zero in this specification. The difference between these coefficients is only marginally significant in column 1 but becomes more precise as controls are added to the regression in columns 2 and 3. Together, the Wald tests presented at the bottom of Table 4 show that when participants disagree with their partner, they are more likely to override a woman than a man.[41]

The estimates presented in columns 2 and 3 of Table 4 control for two important variables that are observable in the photo treatment: partner gender and partner ethnicity. However, there are other characteristics visible in photographs which may be correlated with gender and may affect overriding behavior in the photo treatment. To study whether controlling for these traits affects our estimates of treatment effects and gender differences within the photo treatment, we conducted a second experiment where we collected third-party ratings of all participant pictures taken in the photo treatment. Raters were asked to score photos along 5 dimensions: (1) perceived ability, (2) confidence, (3) attractiveness,

[40] This is the lower bound of the 95% confidence interval of the "Female decision maker" coefficient presented in column 6 of Table 3. The lower bound ranges between 10.35 and 3.29 percentage points across columns 4-6 of Table 3, and between 7.49 and 2.18 percentage points when cases of agreement are included in the estimation sample (columns 1-3 of Appendix Table 5, Panel B).

[41] Estimates of a model that fully interacts partner gender with treatment rather than with the photo treatment only are presented in Appendix Table 6 and shows similar results. In Appendix Table 7 we present mean overriding rates by gender of the decision maker, gender of the partner, and treatment. Both men and women seem to override women more than men in the photo treatment; however, we are underpowered detect these differences in the subsample of male and female decision makers.

(4) the likelihood of being a native English speaker, and (5) someone who is nice.[42] Scores were incentivized, raters earned money if their answer coincided with the most frequent choice made by other subjects.[43] We use the mean score provided by raters to each photo as our trait measure of interest, normalized to have a mean 0 and a standard deviation of 1. Online Appendix E describes the exact procedures used to collect data in this second experiment.

To study whether these ratings and a few other observable characteristics affect overriding behavior and our estimates of gender differences within the photo treatment, we present in Table 5 estimates of a linear probability model of overriding that includes question dummies, standard errors clustered at the group level, and examines one by one the role played by each observable characteristic. Estimates presented in columns 2-10 of Table 5 reveal that few characteristics other than partner gender significantly affect the likelihood that participants override in the photo treatment. The only two exceptions are social distance (column 4) and whether the participant is perceived to be nice (column 10). Column 4 shows that participants who are paired with someone they know, override more than those paired with a stranger. However, there are only 5 participants who report knowing their partner (5 men and 3 women).[44] Column 10, on the other hand, shows that decision makers who are perceived to be nice are less likely to override their partner. This suggests that other-regarding preferences could in part be explaining overriding in the photo treatment. Importantly, neither variable significantly

[42] These 5 traits were chosen because they represent observable features that may affect the likelihood that participants override in the experiment, either through belief updating and/or overriding preferences. For example, higher perceived partner ability and confidence may make participants more likely to agree with someone. Higher partner attractiveness and lower perceived likelihood to be nice, on the other hand, may make it costlier for participants to override. There is a literature documenting a beauty premium (e.g., Biddle and Hamermesh 1998, Mobius and Rosenblat 2006) and differential treatment of participants who are smiling (e.g Scharlemann et al 2001), which would support these hypotheses. We also elicit the likelihood that participants are perceived to be a native English speaker because two non-incentivized questions asked during the post-experiment survey revealed that native English speakers and Australian-born participants are believed to have a relative advantage in the experimental task. Sixty eight percent of participants believe that native English speakers outperform non-native English speakers, while 35% believe that Australian born participants outperform their foreign-born counterparts (51% believe there is no difference in performance by place of birth).

[43] Similar incentives are used by Xiao and Houser (2005) and Grosskopf and Pearce (2020) to elicit the traits of messages or pictures and by Krupka and Weber (2013) to elicit social norms. Studies using non-incentivized image ratings include Biddle and Hamermesh (1998); Mobius and Rosenblat (2006); Wilson and Eckel (2006); Andreoni and Petrie (2008); and Mujcic and Frijters (2020).

[44] We asked participants in the photo treatment three questions at the end of the experiment to measure social distance between partners. These questions asked whether their partner is: (1) someone they know; (2) someone they communicate with on a regular basis; and (3) their friend. We use the answer to the second question to construct our social distance indicator because all participants who claim to know or be friends with their partner, also say they communicate with them on a regular basis. There are only 2 participants, both men, who report being friends with each other in the photo treatment.

affects our estimates of the of the role of decision-maker or partner gender. Across all columns we see no evidence of differences in overriding by gender of the decision maker, and a similarly sized and marginally significant response to being paired with a woman.

Appendix Table E2 in Online Appendix E shows that the lack of correlation between the elicited trait measures and overriding is not driven by poor data quality. The 5 traits we elicited vary with observable photo characteristics such as gender, ethnicity, facial expressions, and even the accessories participants wear. They are also correlated with each other. In Appendix Table 8, we present estimates of Table 5 that control for round and question sequence instead of including question dummies. Estimates of gender differences are more precise. In Appendix Table 9, we simultaneously control for all characteristics. Both tables show evidence in support of the hypothesis that women are overridden more than men in the photo treatment.[45]

In Appendix Table 10 we check whether the results presented in this section are robust to including cases of agreement in the estimation sample and to controlling for the additional observable photo characteristics presented in columns 4-10 of Table 5. While the comparative statics are similar, some precision is lost when we include cases of agreement in the estimation sample. We are underpowered to detect the 4-percentage point difference between the likelihood that participants override men versus women in the photo treatment at the 0.05 level of significance. The p-values from the Wald tests presented in Appendix Table 10 vary between 0.058 and 0.199 depending on the set of controls included in the model.

Overriding observed versus predicted by beliefs (Tasks 2 and 3)

Since men are slightly better at the task than women and this could give rise to the differential treatment we document in the previous subsection, we also estimate a model similar to the one presented in equation (2), which uses Task-2 and 3 data to determine whether beliefs or preferences can explain the differential overriding of men and women. Column 1 of Table 6 shows that participants experience a disutility from overriding men and women in the photo treatment. Although the size of this disutility is larger in magnitude when participants are paired with men than when they are paired with women

[45] There is some evidence of a beauty and perceived ability premium in the joint model presented in Appendix Table 9. Relative attractiveness and perceived ability are associated with a lower likelihood of being overridden.

(30 vs 22 percentage points), we cannot reject the null hypothesis that these intrinsic overriding costs are equal. The overriding rate predicted by beliefs in Task 3 is, nevertheless, equal for partners of both genders suggesting that beliefs do not explain the differential treatment of men and women. Adding controls for individual characteristics has little effect on the coefficients of interest in column 2. Controlling for disagreement size, history of play, and risk preferences, however, does affect our estimates of the size of the gap between observed and predicted overriding rates. It suggests that there is on average a distaste for overriding that is unique to the photo treatment when participants are paired with men, but not necessarily with women. In Appendix Table 11 we present estimates of Table 6 that include cases of agreement in the estimation sample and either exclude or include additional controls for observable partner characteristics. Results are similar.

Belief updating (Task 3)

Since the experiment elicits prior and posterior beliefs in Tasks 1 and 3, we can also examine whether there is heterogeneity in belief updating by gender and treatment. To do this, we transform beliefs to log odds ratios and estimate a linear model of belief updating where the dependent variable is the log odds ratio of the posterior belief that the choice made by Person 1 is correct. The independent variables are the log odds ratio of the prior belief attached to that same choice in Task 1, the gender of the partner in the photo treatment, and the gender of the decision maker.[46] We estimate an interacted model via ordinary least squares, including question fixed effects, and standard errors clustered at the group level. Rather than presenting regression results in tables, we show predictive margins in Figure 6.

Panel A of Figure 6 shows that there are some differences in how men and women update beliefs when participants interact anonymously in the control treatment. On average, women update more than men when their prior is low but not when it is high. The difference in intercept is marginally significant (p=0.056). Panel B and C of Figure 6 show predictive margins when participants are paired with a man and woman in the photo treatment. We cannot reject the null hypotheses of equal updating by gender of the partner or by gender of the decision maker. A comparison of belief updating across

[46] We transform beliefs of 0 and 100 to 0.1 and 99.9 respectively in order to compute odds ratios. See Appendix Figure 11 for a binned scatterplot of the raw data that does not transform beliefs to log odds ratios.

panels within gender of the decision maker, however, shows some marginally significant differences. Men update more when their prior is low and their partner is a woman relative to when the partner is unknown (intercept p=0.066), a difference that disappears for high priors. Belief updating is not different across panels for female decision makers.

These results complement the analysis performed in the previous subsection, which showed that beliefs alone cannot explain overriding behavior or the differential treatment of men and women. For beliefs to explain why we find evidence in support of the hypothesis that women are overridden more than men, we would need to see participants updating their beliefs more when they are paired with men than when they are paired with women.

5.5. Group performance

Having analyzed individual behavior in Tasks 1-3, we proceed with the analysis of group performance. We first ask whether overriding helps groups. To answer this question, we estimate a linear probability model of the group answering the question correctly as a function of whether Person 2 overrides, the gender of Person 1, and the gender of Person 2. We include question fixed effects like in previous models, and cluster standard errors at the group level. Estimates provided in Table 7 show that overriding does not significantly help groups on average in the control treatment; there is an insignificant 7 percentage point improvement in performance. However, overriding marginally improves group performance in the photo treatment, the coefficient of P2 overrides is 13 percentage points but not statistically different than in the control treatment. Pooling data from both treatments, we see that overriding generates an 11-percentage point (27%) improvement in group performance. Table 7 also shows that the gender of Person 1 and Person 2 does not affect the probability that a group answers a question correctly in the experiment, irrespective of whether we control for whether Person 2 overrides. In results not shown we also check whether results are robust to controlling for whether the choice made by Person 1 was correct. Results are similar.[47]

[47] We conduct this robustness check because participants may be differentially likely to answer a question correctly when assigned the role of Person 1 depending on whether their partner is a man or a woman in the photo treatment. An OLS regression of Person 1 choosing the correct answer on a female partner indicator that includes question dummies and clusters standard errors at the group level reveals statistically insignificant gender differences in the photo treatment. The coefficient on female partner is 0.046 (s.e. 0.728 p=0.527) among male

Estimates presented in Table 7 suggest that inefficient overriding might be deterred in the photo treatment. To formally examine this, we use Task-3 data to compare counterfactual group performance in Task 3 to observed group performance in Task 2, as a function of observed versus predicted overriding.[48] Results presented in Appendix Table 12 show that the additional overriding that would occur if participants did not experience a disutility from overriding would not significantly affect group performance in either treatment. The coefficient of Predicted X Override T2&3 is, nevertheless, positive in the control treatment and negative in the photo treatment, suggesting that social image concerns might be deterring inefficient overriding in the photo treatment.

5.6 Robustness checks

We perform a series of robustness checks. First, as already discussed in the previous subsections, we check whether including cases of agreement in the estimation sample changes any of the conclusions drawn. Because overriding should not occur when participants agree with each other, unless participants experience a joy from overriding, which we find no evidence of in the data, some precision is lost but the comparative statics are similar.[49] Second, to study behavior across rounds rather than within question we also estimate linear probability models of overriding that include controls for round and a question sequence rather than question dummies. Results are similar and sometimes even more precise—see the appendix tables when models are not directly presented in the main tables. Since our design cannot tease apart question from round effects, we do not lead with these specifications. Third, in results not shown, we redefine overriding in Task 2 as resubmitting an answer and not necessarily a different answer. This no longer captures voicing disagreement but allows us to include resubmissions of the same answer in the analysis, which occur 6.45% of the time.[50] Estimates of treatment effects and differences in overriding behavior by gender of the decision maker in Task 2 are

decision makers, 0.056 (s.e. 0.065, p=0.390) among female decision makers, and 0.059 (s.e. 0.044, p=0.182) in the pooled sample of men and women.

[48] The formal model that we estimate using ordinary least squares is: $Group\_answer\_correct\_t2\&3_{igq} = \beta_0 + \beta_1 Override\_t2\&3 + \beta_2 Predicted + \beta_3 PredictedXOverridet2\&3 + \gamma_q + \epsilon_{ig}$. Variable $Override\_t2\&3$ is equal to override in Task 2 and predicted override in Task 3. The dependent variable is a dummy variable equal to 1 when the group answer is correct in Task 2 and when the predicted group answer is correct in Task 3.

[49] Participants override their partner when they agree with them 13 times in the data. These cases represent 3.67% of all agreement cases, 1.75% of all overriding opportunities, and do not occur at significantly different rates across treatments.

[50] Resubmissions of the same answer occur at similar rates across treatments (Mean control: 5.56%, Mean photo: 6.96%, p=0.541).

similar.[51] Fourth, we check whether including data from the groups where at least one participant does not self-identify as male or female changes our results. Estimates of models that include the full sample of observations and a female dummy for own and partner gender provide similar results. Finally, we conduct a series of robustness checks of the data collected in the third-party photo ratings experiment, which are discussed in Online Appendix E and do not affect any of the conclusions we draw.[52]

6. Conclusion

We conduct a laboratory experiment to study the factors that affect individual willingness to speak up after someone has raised an opinion. Participants answer multiple-choice general knowledge questions in pairs. One team member is randomly selected to contribute an answer first in each question. The other person sees the choice made by their partner and chooses whether to resubmit an answer on behalf of the group. The second person therefore has full authority over the group decision. We call the resubmission of a different answer overriding and study this behavior in two treatments that vary whether interactions occur anonymously. We compare behavior in a photo treatment, where participants see the photo of the person they are paired with, to a control treatment, where interactions occur anonymously. Using a series of tasks, we also elicit incentivized beliefs about the likelihood that each possible answer option to a question is correct. This allows us to identify the extent to which disagreement explains overriding and to identify the role of observability and social image.

Results show that participants are less likely to override ideas when interactions are not anonymous. This treatment effect is explained by social image costs. We cannot reject the null hypothesis that belief updating fully explains overriding behavior in the anonymous control treatment but find a statistically significant gap between the rate of overriding observed versus predicted by beliefs in the photo treatment. We show that part of this gap is unique to the photo treatment and interpret it as being driven by social image. Examining heterogeneity in behavior by gender reveals that men and

[51] Differences in overriding behavior by gender of the partner reveal the same comparative statics but the coefficients are smaller in size and not statistically significant.

[52] We conduct two types of robustness checks of the third-party photo ratings data which aim to exclude noisy ratings from the construction of the mean trait rating per photo. One identifies participants who exhibit little variation in ratings and excludes them from the construction of the trait variable. The other excludes fast ratings, which are error-prone (see e.g., Recalde et al. 2018). Results are robust to using a restricted sample of ratings to construct the mean trait rating per photo.

women do not override at significantly different rates. Contrary to our hypothesized prediction, we can rule out women overriding less than men in the photo treatment. We also provide some evidence that men and women are treated differently. When participants disagree with their partner, they are more likely to override a woman than a man. This differential treatment is not explained by other observable partner characteristics that are correlated with gender and seems to be in part explained by preferences.

This paper contributes to the literature by studying willingness to speak up and voice disagreement in teams after someone has raised an opinion. Previous work has examined who speaks up first in groups, and group deliberations in environments where who speaks up when is endogenous. We study an environment where individuals can override others and identify the channels underlying this decision. Since speaking up often involves publicly (dis)agreeing with someone, or with the status quo, and a hesitation to disagree with others may affect group performance, this study sheds light on the factors that affect team deliberations and performance.

By studying the role of observability and social image, the study also has a few implications for policy design. First, since individuals are more willing to voice disagreement under anonymity than when their identity is revealed, organizations seeking to encourage people to speak up and voice disagreement could consider eliciting ideas in anonymous environments or providing an option for employees to express their disagreement anonymously in some scenarios. Second, we find that overriding increases with repeated interactions and is higher when the social distance between team members is lower. This suggests that fostering inclusive team relationships which reduce the social distance between team members could increase the likelihood that employees speak up and voice disagreement. More work is needed to study these conjectures.

Third, within the context of gender, our study shows that women are not more hesitant than men to override when they have full authority to do so. This alleviates the concern some people might have regarding the possibility that policies seeking to promote the advancement of women in organizations and in society may backfire if women may hesitate to make difficult choices when they are in positions of power. We find no evidence that giving women rather than men override authority affects groups performance. Since our sample is diverse, reflecting the characteristics of the subject

pool we use, it would be interesting to study the robustness of this result in other study samples and environments.

More work is needed to understand the factors that give rise to the gender gap in speaking time and in communication patterns observed in professional environments where women are underrepresented. Our finding that women do not override less than men in the photo treatment suggests that environmental factors related to who speaks up first and who is heard (or expects to be heard) in a group may matter as well as gender differences in who experiences backlash for speaking up. Future work could study these and other factors that likely play a role such as audience and reputation effects, the stereotype associated with the task, and the gender composition of teams. Since some of these factors affect who speaks up first in groups (e.g., Coffman 2014, Chen and Houser 2019), they may affect overriding behavior and help explain the overall gender gap in speaking time documented in the literature. It is, of course, also possible that our results are driven by the specific aspects of our study design and sample. Studying the robustness of our result together with the role of other factors seems like a fruitful avenue for future research.

Beyond gender, there are other factors that may affect overriding behavior in teams such as race and status. The empirical literature on health and patient-physician-race matches, for example, shows that communication, patient satisfaction, and health outcomes improve when Black patients are matched with Black physicians (e.g., Cooper et al. 2003, Alsan et al. 2018, Hill et al. 2020). Since doctors make choices after hearing from patients in environments where they have full authority over diagnosis and prescribed care, this resembles the type of scenario we study.

The same type of dynamics also arises in teams when leaders must aggregate ideas and make choices on behalf of the group. Born et al. (2020) shows in a laboratory experiment that male team members have more influence over the leader's decision than female team members. This is consistent with our findings. What the literature has yet to do is to study the effect of status on group deliberations and overriding behavior. While our study takes a first step in this direction by exogenously varying who in a group has override authority, power is equal across players over rounds in our experiment. Future work could study overriding behavior and group deliberations in environments where override authority is earned or where there are other forms of asymmetries in power.

References

Alsan, M., Garrick, O., & Graziani, G. (2019). Does diversity matter for health? Experimental evidence from Oakland. *American Economic Review*, 109(12), 4071-4111.

Amanatullah, E. T., & Morris, M. W. (2010). Negotiating gender roles: Gender differences in assertive negotiating are mediated by women's fear of backlash and attenuated when negotiating on behalf of others. *Journal of Personality and Social Psychology*, 98(2), 256.

Andreoni, J. and Bernheim, B. (2009). Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects. *Econometrica*, 77(5), 1607-1636.

Andreoni, J. and Petrie, R. (2004). Public goods experiments without confidentiality: a glimpse into fund-raising. *Journal of Public Economics*, 88(7-8), 1605-1623.

Andreoni, J., & Petrie, R. (2008). Beauty, gender and stereotypes: Evidence from laboratory experiments. *Journal of Economic Psychology*, 29(1), 73-93.

Andreoni, J., Rao, J. and Trachtman, H. (2017). Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving. *Journal of Political Economy*, 125(3), 625-653.

Babcock, L., & Laschever, S. (2003). Women don't ask: Negotiation and the gender divide. *Princeton University Press*.

Babcock, L., Recalde, M. P., Vesterlund, L., & Weingart, L. (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107(3), 714-47.

Bertrand, M. (2011). New perspectives on gender. In Card, D., & Ashenfelter, O. (Eds.) *Handbook of Labor Economics* (Vol. 4, pp. 1543-1590). Elsevier.Bertrand, M. (2018). Coase Lecture–The Glass Ceiling. *Economica*, 85(338), 205-231.

Biddle, J. E., & Hamermesh, D. S. (1998). Beauty, productivity, and discrimination: Lawyers' looks and lucre. *Journal of labor Economics*, 16(1), 172-201.

Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3), 789-865.

Bohnet, I., & Frey, B. (1999). The sound of silence in prisoner's dilemma and dictator games. *Journal Of Economic Behavior & Organization*, 38(1), 43-57.

Bohnet, I., Van Geen, A., & Bazerman, M. (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, 62(5), 1225-1234.

Bohren, J. A., Imas, A., & Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American Economic Review*, *109*(10), 3395-3436.

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2019). Beliefs about Gender. *American Economic Review*, 109(3), 739-73.

Born, A., Ranehill, E., & Sandberg, A. (2020). Gender and willingness to lead: Does the gender composition of teams matter?. *The Review of Economics and Statistics*, 1-46.

Bowles, H. R., Babcock, L., & Lai, L. (2007). Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask. *Organizational Behavior and Human Decision Processes*, 103(1), 84-103.

Brescoll, V. L. (2011). Who Takes the Floor and Why: Gender, Power, and Volubility in Organizations. *Administrative Science Quarterly*, 56(4), 622-641.

Burfurd, I., & Wilkening, T. (2018). Experimental guidance for eliciting beliefs with the Stochastic Becker–DeGroot–Marschak mechanism. *Journal of the Economic Science Association*, 4(1), 15-28.

Bursztyn, L., Fujiwara & Amanda Pallais. 2017. "'Acting Wife': Marriage Market Incentives and Labor Market Investments." *American Economic Review*, 107(11), 3288-3319.

Bursztyn, L., & Jensen, R. (2017). Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure. *Annual Review Of Economics*, 9(1), 131-153.

Buser, T., Gerhards, L., & Van der Weele, L. (2018). Measuring responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty*, 56(2), 65–192.

Carli, L. L. (1990). Gender, language, and influence. *Journal of Personality and Social Psychology*, 59(5), 941.

Carter, A., Croft, A., Lukas, D., & Sandstrom, G. (2018). Women's visibility in academic seminars: Women ask fewer questions than men. *PLoS ONE*, 13(9).

Charness, G., Cobo-Reyes, R., Meraglia, S., & Sánchez, Á. (2020). Anticipated discrimination, choices, and performance: Experimental evidence. *European Economic Review*, 127, 103473.

Chen, J., & Houser, D. (2019). When are women willing to lead? The effect of team gender composition and gendered tasks. *The Leadership Quarterly*, 30(6), 101340.

Coffman, K. B. (2014). Evidence on Self-Stereotyping and the Contribution of Ideas. *The Quarterly Journal of Economics*, 129(4), 1625–1660.

Coffman, K . B., Collis, M., & Kulkarni, L. (2019). Stereotypes and Belief Updating. Harvard Business School Working Paper.

Coffman, K., Flikkema, C. B., & Shurchkov, O. (2021). Gender Stereotypes in Deliberation and team decisions. *Games and Economic Behavior*, 129, 329 -349.

Cooper, L. A., Roter, D. L., Johnson, R. L., Ford, D. E., Steinwachs, D. M., & Powe, N. R. (2003). Patient-centered communication, ratings of care, and concordance of patient and physician race. *Annals of Internal Medicine*, 139(11), 907-915.

Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics*, *22*(2), 369-395.

Crosetto, P., & Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, 19(3), 613-641.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2), 448-74.

Danz, D., Vesterlund, L., & Wilson, A. J. (2020). Belief elicitation: Limiting truth telling with information on incentives (No. w27327). National Bureau of Economic Research.

Dellavigna, S., List, J., Malmendier, U., & Rao, G. (2016). Voting to Tell Others. *The Review Of Economic Studies*, 84(1), 143-181.

Dupas, P., Modestino, A. S., Niederle, M., Wolfers, J. & The Seminar Dynamics Collective (2021). Gender and the dynamics of economics seminars (No. w28494). National Bureau of Economic Research.

Eckel, C. C., & Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results*, 1, 1061-1073.

Eckel, C., & Petrie, R. (2011). Face Value. *American Economic Review*, 101(4), 1497-1513.

Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, 80(3), 532–545.

Exley, C. L., & Kessler, J. B. (Forthcoming). The gender gap in self-promotion. *The Quarterly Journal of Economic.*

Filippin, A., & Crosetto, P. (2016). A reconsideration of gender differences in risk attitudes. *Management Science*, 62(11), 3138-3160.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics, Springer; Economic Science Association*, 10(2), 171-178.

Funk, P. (2010). Social Incentives and Voter Turnout: Evidence from the Swiss Mail Ballot System. *Journal Of The European Economic Association*, 8(5), 1077-1103.

Gago, A. (2018). Confrontation Costs in Negotiations: Bargaining Under the Veil of a Screen. Working paper.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal Economics Science Association* 1, 114–125.

Grosskopf, B., & Pearce, G. (2020). Do You Mind Me Paying Less? Measuring Other-Regarding Preferences in the Market for Taxis. *Management Science*, 66(11), 5059-5074.

He, S., Offerman, T., & van de Ven, J. (2017). The Sources of the Communication Gap. *Management Science*, 63(9), 2832-2846.

Hill, A., Jones, D., & Woodworth, L. (2020). Physician-Patient Race-Match Reduces Patient Mortality. Working paper.

Holt, C. A., & Laury, S. K. (2002). Risk Aversion and Incentive Effects. *American Economic Review*, 92(5), 1644-1655.

Isaksson, S. (2018). It Takes Two: Gender Differences in Group Work. Working Paper.

Jacobi, T., & Schweers, D. (2017). Justice, interrupted: The effect of gender, ideology, and seniority at Supreme Court oral arguments. *Virginia Law Review*, 1379-1485.

Jones, D., & Linardi, S. (2014). Wallflowers: Experimental evidence of an aversion to standing out. *Management Science,* 60(7), 1757-1771.

Karni, E. (2009). A Mechanism for Eliciting Probabilities. *Econometrica,* 77(2), 603-606

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary?. *Journal of the European Economic Association*, 11(3), 495-524.

Manian, S., & Sheth, K. (2021). Follow my lead: Assertive cheap talk and the gender gap. *Management Science*.

Mas, A., & Moretti, E. (2009). Peers at Work. *American Economic Review*, 99(1), 112-145.

Mobius, M. M., Niederle, T. & Niehaus, P., & Rosenblat, T. S. (Forthcoming). Managing self-confidence: Theory and experimental evidence. *Management Science*.

Mujcic, R., & Frijters, P. (2021). The Colour of a Free Ride. *The Economic Journal*, 131(634), 970-999.

Niederle, M (2016) Gender. In Kagel, J. H., & Roth, A. E. (Eds.). *The Handbook of Experimental Economics* (Vol. 2). Princeton University Press.

Niederle, M., & Vesterlund, L. (2007). Do Women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 122(3), 1067-1101.

Recalde, M., & Vesterlund, L. (2020). Gender differences in negotiation and policy for improvement (No. w28183). National Bureau of Economic Research.

Recalde, M. P., Riedl, A., & Vesterlund, L. (2018). Error-prone inference from response time: The case of intuitive generosity in public-good games. *Journal of Public Economics*, 160, 132-147.

Rudman, L. A. (1998). Self-promotion as a risk factor for women: the costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology*, 74(3), 629.

Rudman, L. A., & Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in Organizational Behavior*, 28(6-79).

Sarsons, H., (2019) Interpreting Signals in the Labor Market: Evidence from Medical Referrals. Working paper.

Sarsons, H., Gërxhani, K., Reuben, E., & Schram, A. (2021). Gender differences in recognition for group work. *Journal of Political Economy*, 129(1), 101-147.

Scharlemann, J. P., Eckel, C. C., Kacelnik, A., & Wilson, R. K. (2001). The value of a smile: Game theory with a human face. *Journal of Economic Psychology*, 22(5), 617-640.

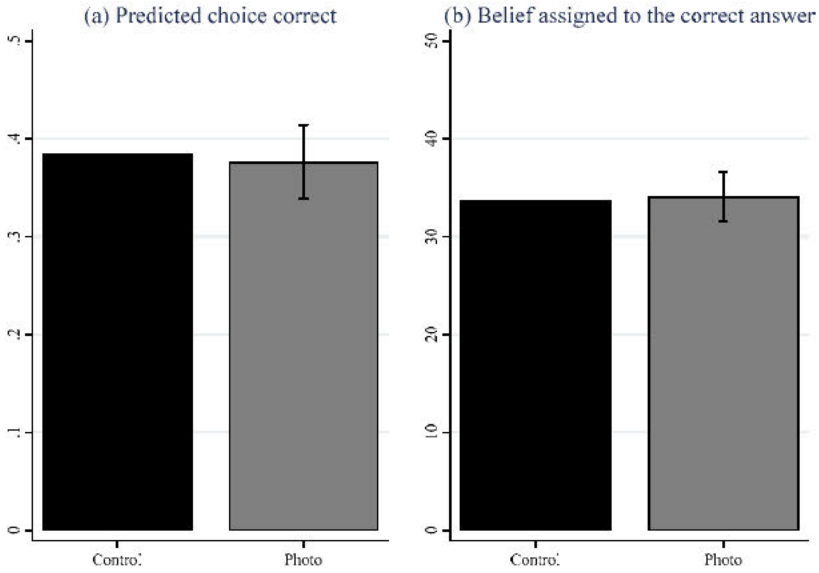
Smith-Lovin, L. & Brody, C. (1989). Interruptions in Group Discussions: The Effects of Gender and Group Composition. *American Sociological Review*, 54(3), 424-435

Snyder, K. (2014). Want to get ahead as a woman in tech? *Language Log*. U of Pennsylvania. https://languagelog.ldc.upenn.edu/nll/?p=13513

Wilson, R. K., & Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly*, 59(2), 189-202.

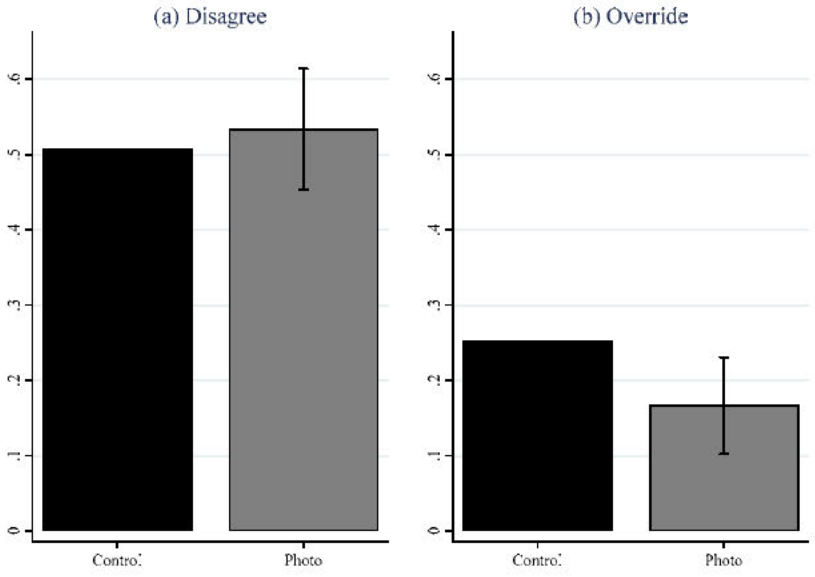
Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, 102(20), 7398-7401.

Figure 1. Mean ability by treatment, Task 1



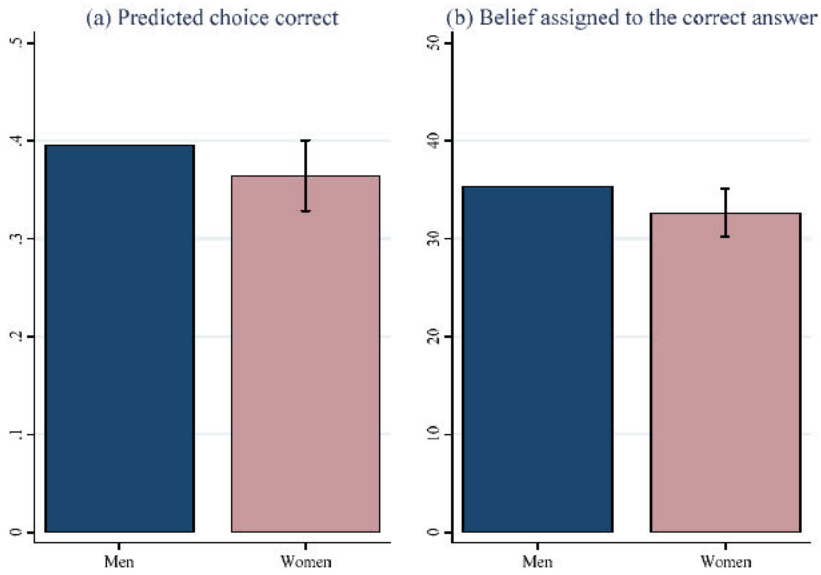
Note: Individual ability is equal to the proportion of predicted choices that are correct in Panel A and to the mean belief assigned to the correct answer across all questions in Panel B. Predicted choices are the answer option(s) that a participant assigns the highest probability weight to in any given question. Whiskers show 95 percent confidence intervals calculated from OLS regressions of individual ability on an indicator for being in the photo treatment.

Figure 2. Disagreement and overriding rates by treatment, Task 2



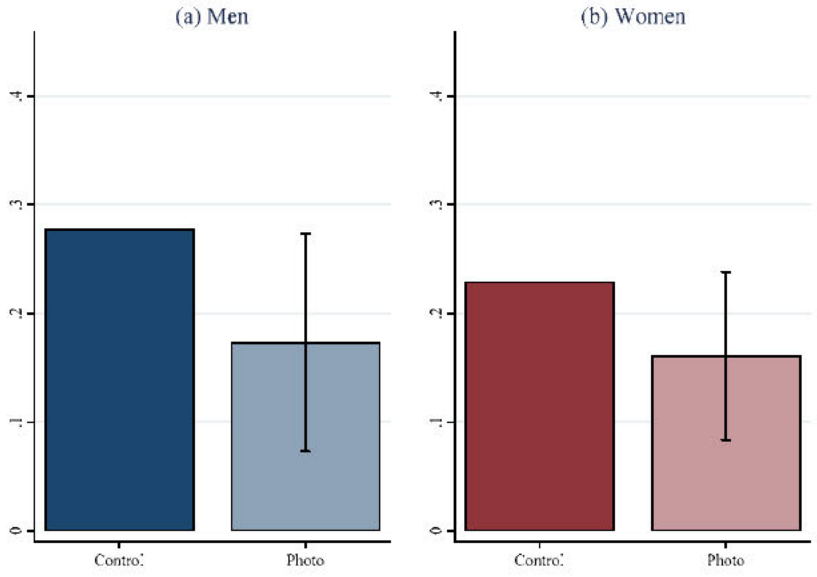
Note: Whiskers show 95 percent confidence intervals calculated from OLS regressions of the dependent variable on an indicator for being in the photo treatment. Standard errors are clustered at the group level.

Figure 3. Mean ability by gender, Task 1



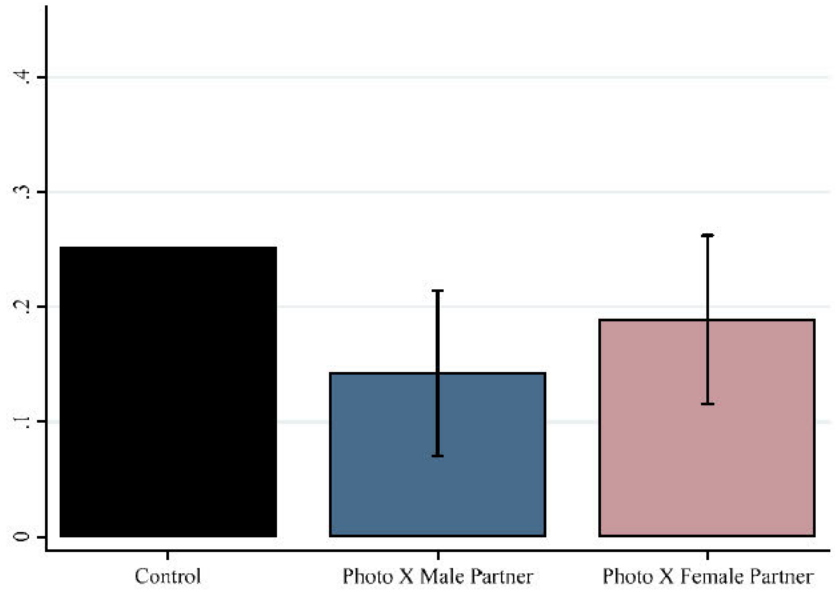
Note: Individual ability is equal to the proportion of predicted choices that are correct in Panel A and to the mean belief assigned to the correct answer across all questions in Panel B. Predicted choices are the answer option(s) that a participant assigns the highest probability weight to in any given question. Whiskers show 95 percent confidence intervals calculated from OLS regressions of individual ability on a female gender indicator. We pool men and women across treatments.

Figure 4. Overriding rate by gender of the decision maker and treatment



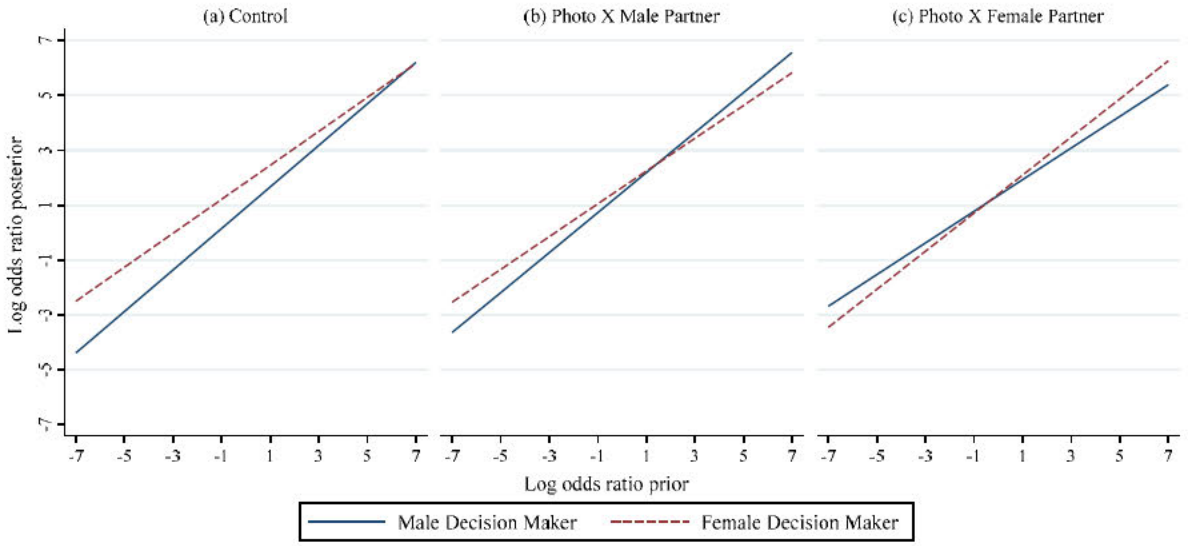
Note: Whiskers show 95 percent confidence intervals calculated from OLS regressions of overriding on a female gender indicator within treatment. Standard errors are clustered at the group level.

Figure 5. Overriding rate by gender of the partner and treatment



Note: Whiskers show 95 percent confidence intervals calculated from OLS regressions of overriding on the interaction of treatment and gender of the partner. Standard errors are clustered at the group level. Appendix Figure 10 distinguishes between male and female partners in the control treatment.

Figure 6. Belief updating by gender



Note: Predictive margins of an OLS regression of posterior beliefs on prior beliefs interacted with gender of the decision maker, gender of the partner, and treatment. Appendix Figure 11 shows binned scatterplots of the raw data that do not transform beliefs to log odds ratios.

Table 1. Treatment effects

Dep. Var.: Override T2	(1)	(2)	(3)	(4)	(5)	(6)
Photo	-0.1367***	-0.1500***	-0.1432***	-0.1335***	-0.1349**	-0.1323**
	(0.0503)	(0.0538)	(0.0523)	(0.0507)	(0.0549)	(0.0525)
Ability score		0.0087***	0.0093***		0.0090***	0.0094***
		(0.0024)	(0.0026)		(0.0023)	(0.0025)
Disagreement size			0.0048***			0.0052***
			(0.0006)			(0.0006)
Overridden before			-0.0168			-0.0211
			(0.0565)			(0.0561)
Risk seeking			0.0070			0.0050
			(0.0096)			(0.0099)
Round				0.0494***	0.0508***	0.0300**
				(0.0143)	(0.0139)	(0.0138)
Includes controls for:						
Question (dummies)	Yes	Yes	Yes			
Individual characteristics		Yes	Yes		Yes	Yes
Question sequence				Yes	Yes	Yes

Note: The sample includes cases of disagreement only (390 observations, 124 clusters). The mean of the omitted category is 0.4380. Ability score is the mean belief assigned to the correct answers in Task 1. Disagreement size ranges from 1 to 100 and indicates the difference between the maximum belief assigned to an answer option in Task 1, and the Task-1 belief assigned to the answer option chosen by Person 1 in Task 2. Overridden before is an indicator of whether the participant was overridden at least once in a previous round. Standard errors clustered at the group level shown in parentheses. ***p<0.01, **p<0.05, *p<0.01.

Table 2. Overriding observed versus predicted by beliefs

Dep. Var.: Override T2&3	(1)	(2)	(3)	(4)	(5)	(6)
Predicted	0.0708	0.0578	0.0890	0.0490	0.0354	0.0860
	(0.0611)	(0.0605)	(0.0556)	(0.0701)	(0.0687)	(0.0567)
Predicted X Photo	0.1870**	0.1957***	0.1331*	0.2020**	0.2117**	0.1308*
	(0.0748)	(0.0747)	(0.0696)	(0.0830)	(0.0821)	(0.0707)
Linear combination:						
Pred. + Pred. X Photo	0.2579***	0.2535***	0.2222***	0.2510***	0.2471***	0.2168***
	(0.0423)	(0.0428)	(0.0411)	(0.0443)	(0.0446)	(0.0416)
Includes controls for:						
Question (dummies)	Yes	Yes	Yes			
Individual characteristics		Yes	Yes		Yes	Yes
Dis. size, hist. of play, risk prefs.			Yes			Yes
Round and q. sequence				Yes	Yes	Yes

Note: The dependent variable is equal to observed overriding in Task 2 and to overriding predicted by beliefs in Task 3. The sample includes cases of disagreement only (803 observations, 124 clusters). The mean of the omitted category is 0.4380. Standard errors clustered at the group level shown in parentheses. ***p<0.01, **p<0.05, *p<0.01.

Table 3. Do men and women override at different rates?

Dep. Var.: Override T2	Control			Photo		
	(1)	(2)	(3)	(4)	(5)	(6)
Female decision maker	-0.0326	-0.0230	-0.0871	0.0086	0.0084	0.0629
	(0.0770)	(0.0780)	(0.0808)	(0.0563)	(0.0536)	(0.0482)
Ability score		0.0167***	0.0177***		0.0033	0.0036
		(0.0040)	(0.0037)		(0.0029)	(0.0031)
Disagreement size			0.0049***			0.0053***
			(0.0011)			(0.0007)
Overridden before			0.0017			-0.0201
			(0.1003)			(0.0627)
Risk seeking			-0.0178			0.0274**
			(0.0124)			(0.0115)
Mean omitted category		0.4571			0.3089	
N observations	137	137	137	253	253	253
N clusters	45	45	45	79	79	79
Includes controls for:						
Question (dummies)	Yes	Yes	Yes	Yes	Yes	Yes
Individual characteristics		Yes	Yes		Yes	Yes

Note: The sample includes cases of disagreement only. Ability score is the mean belief assigned to the correct answers in Task 1. Disagreement size ranges from 1 to 100 and indicates the difference between the maximum belief assigned to an answer option in Task 1, and the Task-1 belief assigned to the answer option chosen by Person 1 in Task 2. Overridden before is an indicator of whether the participant was overridden at least once in a previous round. Standard errors clustered at the group level shown in parentheses. ***p<0.01, **p<0.05, *p<0.01.

Table 4. Are men and women treated differently in the photo treatment?

Dep. Var.: Override T2	(1)	(2)	(3)
Photo X male partner	-0.1892***	-0.2433***	-0.2379***
	(0.0539)	(0.0637)	(0.0620)
Photo X female partner	-0.0875	-0.1216*	-0.1057*
	(0.0602)	(0.0663)	(0.0624)
Ability score		0.0093***	0.0096***
		(0.0024)	(0.0025)
Disagreement size			0.0048***
			(0.0006)
Overridden before			-0.0034
			(0.0581)
Risk seeking			0.0049
			(0.0092)
T-test (p-value)			
Ph. X Male P. = Ph. X Female P.	0.0698	0.0348	0.0155
Includes controls for:			
Question (dummies)	Yes	Yes	Yes
Individual characteristics		Yes	Yes
Photo X partner ethnicity		Yes	Yes
Disagreement size, hist. of play, risk prefs.			Yes

Note: The sample includes cases of disagreement only (390 observations, 124 clusters). The mean of the omitted category is 0.4380. Ability score is the mean belief assigned to the correct answers in Task 1. Disagreement size ranges from 1 to 100 and indicates the difference between the maximum belief assigned to an answer option in Task 1, and the Task-1 belief assigned to the answer option chosen by Person 1 in Task 2. Overridden before is an indicator of whether the participant was overridden at least once in a previous round. Standard errors clustered at the group level shown in parentheses. ***p<0.01, **p<0.05, *p<0.01.

Table 5. Do other photo characteristics affect overriding behavior?

Dep. Var.: Override T2	Characteristic=									
	Gender	Self-reported ethnicity	Perceived race/ethnicity	Know partner	Age	Ability	Confidence	Attractiveness	Native English speaker	Nice
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Female DM	0.0180	0.0082	0.0169	0.0246	0.0162	0.0137	0.0009	0.0148	0.0182	0.0292
	(0.0559)	(0.0549)	(0.0562)	(0.0556)	(0.0563)	(0.0548)	(0.0578)	(0.0561)	(0.0548)	(0.0546)
Female partner	0.0953*	0.1074*	0.0937*	0.0956*	0.0926	0.0955*	0.0898	0.0964*	0.0957*	0.0942*
	(0.0551)	(0.0573)	(0.0560)	(0.0540)	(0.0559)	(0.0540)	(0.0597)	(0.0553)	(0.0549)	(0.0545)
DM Caucasian		0.0535	0.0529							
		(0.0742)	(0.0707)							
DM Other ethnicity		-0.0730	-0.0396							
		(0.0777)	(0.0669)							
Partner Caucasian		-0.0245	-0.0407							
		(0.0719)	(0.0715)							
Partner Other ethnicity		0.1064	0.0412							
		(0.0707)	(0.0725)							
DM characteristic				0.4116**	-0.0061	-0.0252	-0.0325	0.0018	0.0014	-0.0439*
				(0.1621)	(0.0295)	(0.0262)	(0.0281)	(0.0279)	(0.0262)	(0.0248)
Partner characteristic					-0.0217	-0.0261	-0.0053	-0.0345	0.0020	0.0135
					(0.0315)	(0.0237)	(0.0259)	(0.0290)	(0.0285)	(0.0245)
R-squared	0.1191	0.1356	0.1247	0.1374	0.1208	0.1261	0.1237	0.1245	0.1191	0.1278
Includes:										
Question dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: The sample restrics attention to cases of disagreement in the photo treatment (253 observations, 79 clusters). DM stands for decision maker. Each column regresses overriding in Task 2 on the indicator(s) described in each column. In column 3 we use race/ethnicity variables coded from the photos. In column 4 we use the DM's report of whether they know or have communicated with their partner prior to the experiment. The traits described in columns 6-10 were collected in a separate experiment described in Online Appendix E. We use the mean rating provided by all third-party raters per photo in this separate experiment as our trait proxy, normalized to have mean of 0 and a standard deviation of 1. We also normalize age, the only other non-binary variable included in the table. Standard errors clustered at the group level are shown in parentheses. *p<0.10, **p<0.05, ***p<0.01.

Table 6. Why are men and women treated differently in the photo treatment?

Dep. Var.: Override T2&3	(1)	(2)	(3)
Predicted	0.0707	0.0578	0.0890
	(0.0611)	(0.0603)	(0.0555)
Predicted X Photo X Male Partner	0.2327***	0.2487***	0.1911**
	(0.0830)	(0.0832)	(0.0762)
Predicted X Photo X Female Partner	0.1446*	0.1450*	0.0783
	(0.0857)	(0.0851)	(0.0796)
Linear combination:			
Pred. + Pred. X M partner X Photo	0.3034***	0.3065***	0.2801***
	(0.0560)	(0.0558)	(0.0519)
Pred. + Pred. X F partner X Photo	0.2152***	0.2027***	0.1673***
	(0.0590)	(0.0598)	(0.0563)
Includes controls for:			
Question (dummies)	Yes	Yes	Yes
Individual characteristics		Yes	Yes
Photo X partner ethnicity		Yes	Yes
Disagreement size, hist. of play, risk prefs.			Yes

Note: The dependent variable is equal to observed overriding in Task 2 and to overriding predicted by beliefs in Task 3. The sample includes cases of disagreement only (803 observations, 124 clusters). The mean of the omitted category is 0.4380. Standard errors clustered at the group level shown in parentheses. ***p<0.01, **p<0.05, *p<0.01.

Table 7. Group performance

Dependent variable:	Control			Photo			Both		
Group answer correct	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
P2 overrides	0.0695	0.0662		0.1237*	0.1291*		0.1052**	0.1068**	
	(0.0765)	(0.0758)		(0.0665)	(0.0665)		(0.0501)	(0.0497)	
P1 female		-0.0174	-0.0165		-0.0697	-0.0640		-0.0494	-0.0470
		(0.0692)	(0.0696)		(0.0449)	(0.0451)		(0.0363)	(0.0364)
P2 female		-0.0277	-0.0330		0.0156	0.0169		0.0023	0.0002
		(0.0677)	(0.0677)		(0.0449)	(0.0447)		(0.0365)	(0.0365)
Photo								-0.0023	-0.0110
								(0.0374)	(0.0376)
Mean omitted category	0.3861	0.3636	0.4103	0.3823	0.3580	0.3750	0.3861	0.3636	0.4103
N observations	270	270	270	474	474	474	744	744	744
N clusters	45	45	45	79	79	79	124	124	124
Includes:									
Question dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: P1 and P2 denote Person 1 and 2 respectively. The sample includes cases of agreement and disagreement. Standard errors clustered at the group level shown in parentheses. *p<0.10, **p<0.05, ***p<0.01