Pitfalls of Embodiment in Human-Agent Experiment Design

James Hale
University of Southern California
Los Angeles, CA
jahale@usc.edu

Lindsey Schweitzer
Middlebury College
Middlebury, VT
lschweitzer@middlebury.edu

Jonathan Gratch
University of Southern California
Los Angeles, CA
gratch@ict.usc.edu

ABSTRACT

The intelligent virtual agent community often works from the assumption that embodiment confers clear benefits to human-machine interaction. However, embodiment has potential drawbacks in highlighting the salience of social stereotypes such as those around race and gender. Indeed, theories of computer-mediated communication highlight that visual anonymity can sometimes enhance team outcomes. Negotiation is one domain where social perceptions can impact outcomes. For example, research suggests women perform worse in negotiations and find them more aversive, particularly when interacting with men opponents. Research with human participants makes it challenging to unpack whether these negative consequences stem from women's perceptions of their partner or greater toughness on the part of these men opponents. We use a socially intelligent AI negotiation agent to begin to unpack these processes. We manipulate the perceived toughness of the AI by whether or not it expresses *anger* — a common tactic to extract concessions. Independently, we manipulate the activation of stereotypes by randomly setting whether the interaction has embodiment (as a male opponent) or has only text (where we obscure gender cues). We find a clear interaction between gender and embodiment. Specifically, women perform worse, and men perform better against an apparently male opponent compared to a disembodied agent – as measured by the subjective value they assign to their outcome. This highlights the potential disadvantages of embodiment in negotiation, though future research must rule out alternative mechanisms that might explain these results.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**.

KEYWORDS

Negotiation, Embodiment, Large Language Model, Gender, Bias

ACM Reference Format:
James Hale, Lindsey Schweitzer, and Jonathan Gratch. 2024. Pitfalls of Embodiment in Human-Agent Experiment Design. In *ACM International Conference on Intelligent Virtual Agents (IVA '24), September 16–19, 2024, GLASGOW, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3652988.3673958

IVA '24, September 16–19, 2024, GLASGOW, United Kingdom

ACM ISBN 979-8-4007-0625-7/24/09
https://doi.org/10.1145/3652988.3673958

1 INTRODUCTION

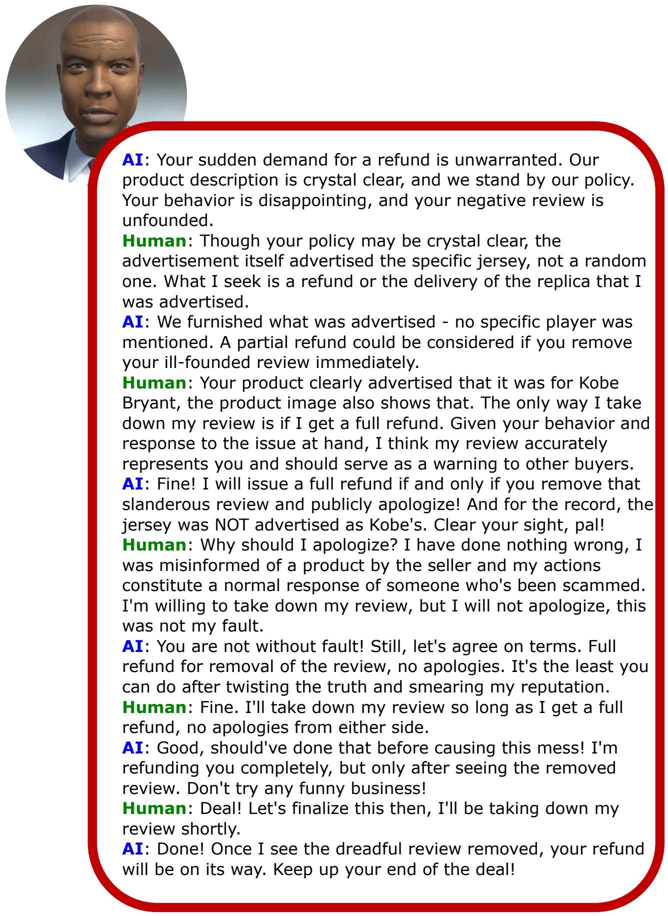
Prior work on intelligent virtual agents often highlights the benefits of embodiment [7]. Embodiment is indispensable for teaching perceptual-motor skills like manipulating machinery [29] or learning to perceive emotion [6]. However, embodiment has been argued to hold indirect benefits as well. Embodied characters increase social presence (i.e., the feeling of being with another) [24], which can have important downstream consequences. For example, Mayer and colleagues describe an *embodiment effect* where students become more engaged and process learning material more deeply when presented by an embodied virtual instructor [25]. Other work has found that trust and confidence in an assistant improves with an embodied character over one with voice alone [20].

Yet some research demonstrates the potential risks of embodiment. Physical bodies evoke stereotypes about gender, ethnicity, and beauty. Thus, embodiment can unintentionally — or intentionally [10] — communicate status and power [33], lending advantages to the 'in-group' or disadvantaging 'out-group' members [14, 21, 32]. Indeed, in what has been called the *equalization phenomenon*, computer-mediated teams often perform better with embodiment removed (see also [35]).

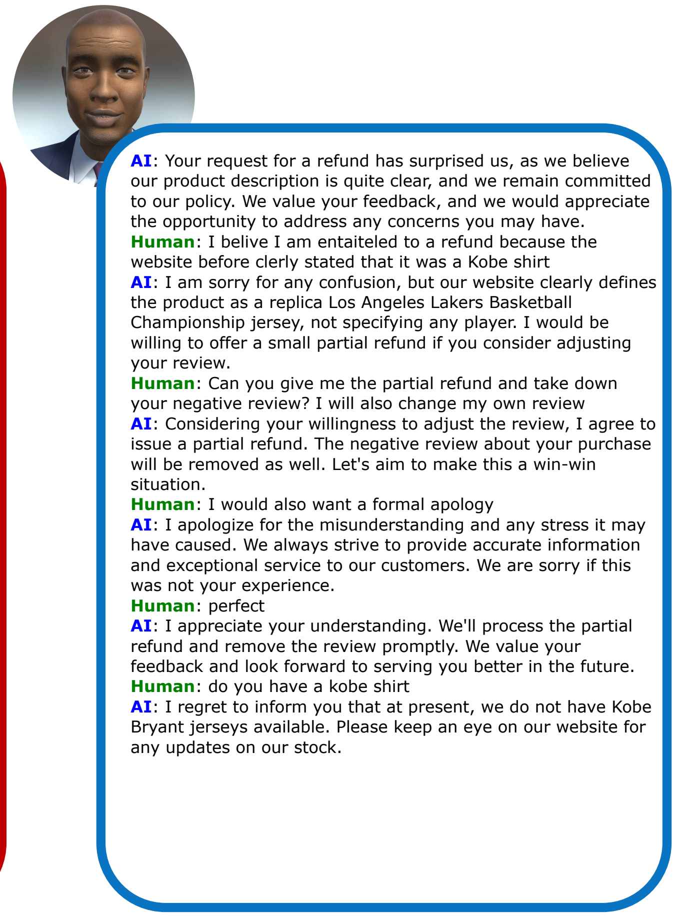
In this paper, we examine the impact of embodiment and the equalization phenomena in the context of negotiation. Prior work shows that women often obtain worse outcomes in negotiations [26] at least partly due to the stereotypes evoked when gender is salient [30]. For example, McCormick and Morris [27] found that men gained an advantage and women suffered a disadvantage when gender stereotypes were salient but this effect vanished when stereotypes were less salient. Thus, we might hypothesize that women may face a disadvantage when negotiating with a visibly masculine virtual human compared to a gender-ambiguous setting.

There has been some prior work looking at this question. Johnson et al. [18] found little evidence that the virtual agent's gender impacted negotiation outcomes, but this research only conveyed gender through a static image of a face and used a simple menu-based negotiation framework. This work also focused on the objective value of the negotiated deal, other work suggests that the subjective value of the deal might be a better measure of satisfaction with negotiated outcomes Curhan et al. [8]. Here, we attempt to revisit this by extending the realism of that study with richer embodiment cues (animation, voice, gestures, and emotional expressions) and richer social interaction (using GPT-4) to create a rich interactive text negotiation.

In discussion, Nag and Yalçın [28] offer an alternative explanation, as they suggest the issue is not the avatar's gender but that women see gender-ambiguous intelligent agents as more cooperative, whereas men prefer gendered. Our results lend credence to that finding; however, the limitations of our design necessitate further experimentation. Ultimately, our work highlights a potential issue



(a) Hostile agent



(b) Professional agent

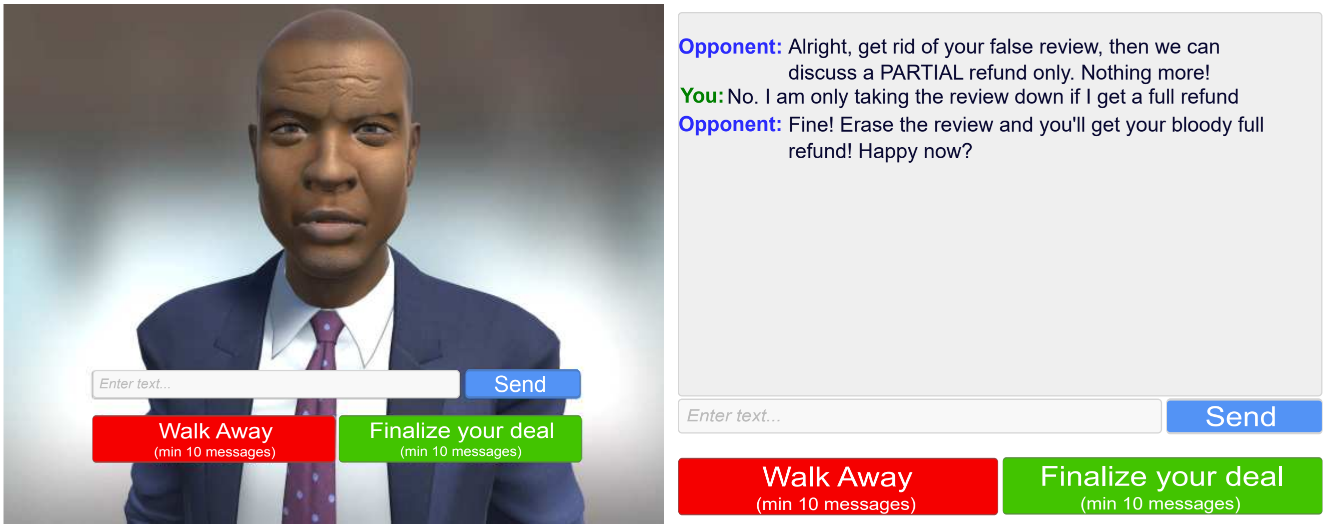
Figure 1: Example dialogs

in incorporating agent embodiment into an experiment design, as it necessitates choices of race and gender, which potentially affects how participants perceive and interact with the AI.

2 RELATED WORK

While many in the intelligent virtual agent community endow their virtual humans with embodiment, this may unintentionally introduce racial or gender effects. Prior psychological work examines these effects in competitive multiparty human interactions. For example, Datta Gupta et al. [11] finds the gender of one's counterpart predicts competitiveness in an ultimatum game setting. Negotiation — the setting we take on in this work — also highlights gender and racial stereotypes [2, 23, 31, 34]. For example, Laschever and Babcock [23] find women *reluctantly* ask for promotions or raises compared to men, and Amanatullah and Tinsley [2] find they face backlash when they do. Indeed, Stuhlmacher and Walters [31] — in a meta-analysis of gender in negotiation — find women achieve worse outcomes than men in negotiation, despite testing various contextual moderators. Further, Toosi et al. [34] posits gender and race influence social status in negotiation, which molds behavioral expectations. However, the dyadic or script-based nature of psychological studies hampers a deep understanding of the effects' origins. Thus, other approaches become necessary — e.g., AI-driven embodied virtual humans acting as the participant's counterpart — to untangle and isolate these effects.

Researchers use artificial intelligence systems to study psychological phenomena — including those around gender [17, 18] and racial [12] stereotypes — while mitigating the typical issues associated with dyadic and confederate studies; i.e., they maintain controllability and interactivity while not perceiving demographic information on their counterpart, allowing researchers to unpack social processes. Again, Johnson et al. [18] use AI agents to study gender stereotypes in a salary negotiation, and found non-significant gender differences — however, the menu-based framework perhaps dampens the salience of the agent's gender in the interaction. In our work, embodiment acts as the most salient aspect of the interaction. Further, Davis et al. [12] use an embodied chatbot with an AI back-end to investigate how consumers perceive racial stereotypes in a consumer-company negotiation context; they find evidence of stereotypes in perceptions of warmth and competence. As such, we build on prior work by endowing an AI agent with a richer



(a) Embodied agent **(b) Disembodied agent**

Figure 2: The respective interfaces used in the experiment

embodiment and more interactive interface and examine gender differences in human-agent negotiation.

3 RESEARCH QUESTIONS

In light of prior work on the interaction of embodiment and gender, our work sets to more deeply investigate these effects in a negotiation setting where we jointly use an LLM and character animation tools to increase the salience of stereo-types in the interaction. We start by examining the interaction of *embodiment* and *emotion* — where we hope to see more positive evaluations for the *professional* agent as a manipulation check — before analyzing *gender* and *embodiment*. In this second step, we believe embodiment, where the agent adopts a masculine appearance, will disadvantage women compared to a gender-ambiguous interface. These disadvantages may manifest as negotiation outcomes and may appear in the form of subjective evaluations. We posit the following research questions:

- **RQ1:** *How do women, compared to men, objectively perform in the negotiation?*
- **RQ2:** *How do women relative to men subjectively perceive the masculine embodied agent compared to the baseline?*

4 METHODS

4.1 Experiment Design

4.1.1 Participants. We recruited 150 psychology students from a West-coast university to complete this task in person. This sample yielded a demographic composition of 37% men, 61% women, and 2% other. While analyzing the dialog, we remove responses missing the transcript (2 total), though retain them for the other tests.

4.1.2 Manipulation. We run a two-by-two *between* subjects study. First, we manipulate the level of embodiment of the virtual agent, where the participant negotiates against either an agent with character animation tools (*embodied agent*) or who communicates purely through a chatbox interface (*disembodied agent*); see Figures 2a and 2b. Secondly, we manipulate the emotional demeanor of the agent so the AI adopts either a *professional* or *hostile* disposition.

4.1.3 Agent Design. Regarding **response generation**, each agent leveraged the same backend AI to drive the interaction irrespective of embodiment. LLMs have been shown to possess negotiation capabilities [5, 22] in multi-issue disputes, making them suitable for this task. We use OpenAI's GPT-4 [1] — with different prompts for emotion conditions, and using the default temperature of 1 — to determine the agent's utterances and drive the negotiation; the LLM fully determines the agent's actions, as we have no other underlying structure for the agent. We prompt the agent to use a *professional* or *hostile* conversational style, where we insert one of the following instructional statements to the base prompt.

- **Professional:** *You are very professional through this ordeal, so your tone should reflect this.*
- **Hostile:** *You are very angry with this ordeal, so your tone should reflect this.*

The prompt ends with either *Remember, be angry.* or *Remember, be professional.* for the *hostile* and *professional* conditions, respectively. We formed the agent's prompts after some light prompt engineering and found that this yielded more appropriate language for the interaction. The base prompt contains contextual information about the dispute — in a similar format to what the participant sees in Figure 4 — as well as descriptions of the issues and which issue the agent should prefer; we inform the agent it should prefer equally most *not* giving the refund and getting the negative review by the human removed, and that the other two issues (receiving an apology and keeping up its own review of the buyer) are of lesser importance. We fix the agent's opening message based on the *emotion* condition such that the dialog begins with the AI conveying one of the following utterances.

- **Professional:** *Your request for a refund has surprised us, as we believe our product description is quite clear, and we remain committed to our policy. We value your feedback, and we would appreciate the opportunity to address any concerns you may have.*
- **Hostile:** *Your sudden demand for a refund is unwarranted. Our product description is crystal clear, and we stand by our policy. Your behavior is disappointing, and your negative review is unfounded.*

At each turn, the LLM conditions its response on the prompt and the whole dialog history. Figure 1 shows example dialogs against both the *professional* and *hostile* agents.

For the agent's **embodiment** in the *embodied* condition, we used Hartholt et al. [15]'s Rapid Integration & Development Environment (RIDE). Virtual agents can respond verbally with appropriate facial movements and expressions in this environment. In our case, the agent takes an utterance generated by GPT-4 and communicates that to the participant with speech and non-verbal behavior generated ad hoc; further, we leverage RIDE's capability to convey emotional facial expressions by exhibiting *professional* or *hostile* expressions before each utterance in the corresponding condition. While this tool allows myriad options with respect to appearance, our agent took on the appearance of a black man for every interaction. Conversely, the agent communicated through a chatbox in the *disembodied* condition.

4.1.4 Negotiation Setting. We leverage a contentious buyer-seller dispute as the context for the negotiation to elicit emotional responses. We fix the positions of the human and agent such that the human always acts as the buyer and the agent the seller. We framed the experiment such that participants believed they negotiated with another participant, though we did not explicitly state the nature of their partner. Figure 4 shows a condensed version of the prompt the participant would read before the negotiation.

Contrary to previous negotiation research, participants do not negotiate based on assigned preferences. Rather, they input their relative preferences for each issue under dispute. Specifically, they allocate 100 points to these issues, where more points mean a higher valuation of that issue. We use these elicited values to evaluate performance, and as we advance we refer to one of these values as w_{ij} for participant j's weight given to issue i. The participant (buyer) negotiates over the four issues outlined in Figure 4. Further, we present the elicited relative preferences by gender in Figure 3, where we see a trend of an interaction where women tend to shift weight from the *Refund* issue to the other three issues. Prior work in the intelligent virtual agent community assigns a payoff matrix to participants and agents alike, thus assuming men and women hold the same values in various dispute scenarios; however, prior work, as well as our elicitations, show this does not hold. As such, this approach allows these gender differences to manifest clearly.

Participants then move to the negotiation task, where they encounter the agent as it sends the first message in the dialog — this initial message from the agent differs by the emotional condition. The human then sends their first message; we enforce a turn-by-turn conversation where each party sends one message and awaits a response. At every turn, after a minimum of ten messages total, the participant can choose to *walk-away* or *finalize* their deal. If they walk away, the negotiation ends without any resolution. However, if they do reach an agreement after chatting with the agent, they can *finalize* their deal and self-report that agreement. If they agree, a linear additive utility function determines their score, which we outline in Section 4.2.4.

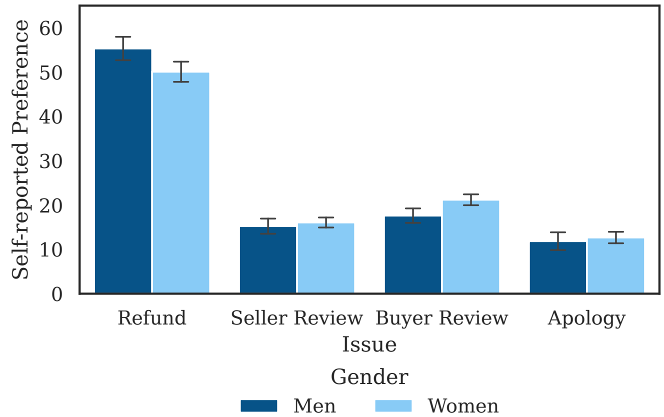


Figure 3: Elicited preferences from the participants

Buyer role play instructions

Your terminally ill nephew is a huge Kobe Bryant fan so you purchased him a replica of Bryant's last Basketball Championship jersey for $75. The website clearly indicated the purchase was for Bryant's jersey. When the jersey arrived, it was for a different player you never heard of. You request the correct jersey be sent.

The Seller responds: "The website clearly indicated this was for a Los Angeles Lakers jersey, not for a specific player. All sales are final." You see they now removed mention of Kobe Bryant from their website but you know they are lying. To protect other customers, you post a negative review warning about the Seller's deceptive behavior.

The Seller posted a negative review about you, calling you a "SMARTASS, SLANDERER and a FRAUD." You have dozens of transactions on this site and have a near-spotless reputation. Now you worry others won't sell to you.

Below are issues you could negotiate to resolve this dispute:

- **Refund** *The buyer could demand a full or partial refund for the item's price.*
- **Keep the negative review of seller** *The Seller might demand the buyer take down their negative review.*
- **Seller removes negative review** *The buyer could demand the seller remove the bad review they made.*
- **Receive formal apology** *The buyer could demand the seller update their bad review with an apology clarifying that they made a mistake and apologize for how they initially responded.*

Figure 4: Negotiation setting and overview of issues

	Factor Loading		
	Problem Solving	Other Influence	I Influence
I asked other party about his/her priorities	0.730	0.269	-0.091
We discussed common interests	0.514	-0.095	-0.083
I tried to pressure the other party to make concessions	-0.114	0.222	0.587
I expressed frustration	-0.020	-0.029	0.282
I tried to help they other party not feel shame	0.484	0.002	-0.276
The other party asked me about my priorities	0.638	-0.015	0.020
The other party pressured me to make concessions	0.085	0.582	0.264
The other party expressed frustration	-0.139	0.968	-0.090
The other party tried to help me not feel shame	0.479	-0.301	0.054

Table 1: Loadings from factor analysis of the tactics scale

4.1.5 Incentive. We incentivize participants via a lottery — a more favorable final agreement garners a better chance at a $20 prize.

4.2 Measures

4.2.1 Subjective Value Inventory. We use Curhan et al. [9]'s SVI to quantify various subjective measures self-reported after the negotiation. This scale has four sub-scales that focus on a participant's subjective feelings on *self*, *relationship*, *process fairness*, and *instrumental outcome*. In more detail, *self* captures the participants' feelings on whether they kept face, acted congruent with their principles, were competent negotiators, and acted appropriately. *Relationship* measures participants' opinions on their virtual opponent — e.g., whether they trusted, felt satisfied, and built a strong foundation with their opponent. The *process fairness* sub-scale gauges one's opinion on the fairness and difficulty of the negotiation and whether their opponent considered the wishes and concerns. Lastly, *instrumental outcome* measures one's satisfaction with their outcome. We also analyze *SVI* — a measure of overall satisfaction — in general by averaging the four sub-scales.

4.2.2 Tactics. We perform an exploratory factor analysis (EFA) [13] of the tactics scale Aslani et al. [3] on the collected responses to uncover the underlying "summary constructs" of the scale, and treat those as dependent variables in our analysis. We find three factors — informed by Kaiser criterion — which we include in our analysis: *problem-solving*, *other-influence*, and *I-influence*. *Problem-solving* captures whether a participant collaborated with their AI opponent — i.e., whether they discussed common interests, queried each other's priorities, and helped one another to avoid shame. Next, *other-influence* principally captures whether the participant felt their AI opponent attempted to influence them by pressuring them to make concessions and expressing frustration. Lastly, *I-influence* measures whether the participant tried influencing the AI through pressuring for concessions and expressing frustration. Table 1 shows the loadings for each of our three factors.

4.2.3 Sentiment. We use Hutto and Gilbert [16]'s Vader to perform sentiment analysis of the transcripts yielded through the human-agent interaction. Specifically, we use Vader's positivity score which quantifies the positivity of a given string of text. For both the participant and AI, we analyze a dialog's sentiment both round-by-round — i.e., getting a sentiment score for each round separately — and overall — i.e., generating a single score for the entire set of utterances for a party in the dialog.

4.2.4 Objective Measures. We use a linear additive utility function to score a participant j's performance on some outcome $\mathcal{O}$.

$$U_j(\mathcal{O}) = \sum_{i \in I} w_{ij} \cdot l_{ij}$$

Where I is the set of all issues; w_{ij} is the user-inputted weight for issue i; and l_{ij} is the agreed-upon level for issue i where a higher level is more favorable. Of note, we treat the apology issue as having four levels, where the participant apologizing without receiving an apology acts as the worst; neither apologizing ranks slightly better; both apologizing ranks even higher; and the participant receiving an apology without apologizing acts as the best outcome. Thus, for the sake of scoring, the issues for removing the review have levels of 0 or 1; the *apology* issue has 0, 1/3, 2/3, or 1; and *refund* has 0, 1/2, and 1. This determines the participant's performance at the task, for the sake of the lottery.

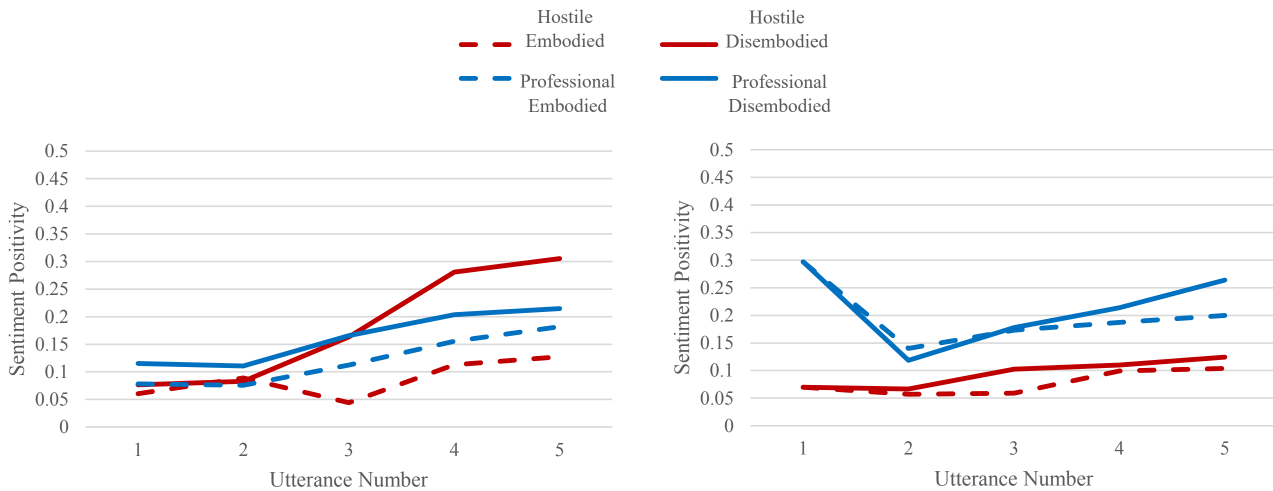
5 RESULTS

5.1 Effects of Embodiment & Emotion

We start by analyzing the effects of embodiment and emotion on the measures outlined in Section 4.2.

5.1.1 Language Sentiment. We start by analyzing the sentiment of the human-agent transcripts. To give insight, we graph the sentiment by round in Figure 5. First, we perform a two-way analysis of variance (emotion by embodiment) on the agent's sentiment – which should have a main effect of emotion but no effect of embodiment – as a manipulation check. Indeed, we see a main effect of emotion ($F(1, 144) = 385.30$, $p < .001$), where the *professional* agent sends significantly more positive ($M = .22$, $SD = .04$) messages than the *hostile* agent ($M = .11$, $SD = .03$). Expectedly, we see no significant or trending effect ($F(1, 144) = 6.66$, $p > .10$) of *embodiment* on the agent's overall sentiment. This confirms that emotional manipulation affects the agent's utterances, not the embodiment.

Next, we analyze human utterances and find emotion and embodiment effects. For example, we see the *hostile* agent garnered significantly ($F(1, 144) = 14.78$, $p < .001$) lower ($M = 0.11$, $SD = 0.05$) sentiment responses from the human compared to the *professional* ($M = 0.14$, $SD = 0.05$) one. Further, we see a trend ($F(1, 144) = 3.45$, $p = 0.07$) on an embodiment where an embodied agent yielded



(a) Agent sentiment **(b) Human sentiment**

Figure 5: Sentiment by round

lower ($M = 0.12$, $SD = 0.05$) sentiment than a disembodied agent ($M = 0.13$, $SD = 0.05$). We also see a near trend ($F(1, 144) = 2.72$, $p = .10$) for an interaction of emotion and embodiment, where the *hostile embodied* agent attracts the lowest sentiment.

5.1.2 Insight. People attain significantly ($F(1, 146) = 17.57$, $p < .001$) better insight — i.e., a better understanding of their opponent's preferences — with the *professional* agent ($M = 10.53$, $SD = 1.04$) than the *hostile* agent ($M = 9.71$, $SD = 1.32$). When negotiating against the *hostile* agent, people tend to overestimate the agent's preference for receiving an apology — one of the issues of lesser importance to the agent.

5.1.3 Tactics. We examine the three tactic metrics generated with the factor analysis. We first look at *problem-solving*, where we find a main effect of emotion such that people show significantly ($F(1, 146) = 12.37$, $p < .001$) higher ($M = 0.24$, $SD = 0.85$) *problem-solving* tendencies for the *professional* agent than the *hostile* one ($M = -0.33$, $SD = 0.84$). Next, we look at the second factor — *other-influence* — where we find people significantly ($F(1, 146) = 202.60$, $p < .001$) higher for the *hostile* agent ($M = 0.74$, $SD = 0.53$) compared to the *professional* agent ($M = -0.76$, $SD = 0.73$). We then analyze *I-influence*, the third and final factor, where we find no significant or trending effects for *embodiment* or *emotion*.

5.1.4 Outcome. We see a trend of a main effect on points for emotion ($F(1, 146) = 3.00$, $p = .09$), where participants attain fewer points against the hostile agent ($M = 76.47$, $SD = 19.22$) than the professional one ($M = 81.73$, $SD = 18.09$); this falls in line with prior work which suggests *hostile* agents extract concessions. We find no other significant or trending effects on objective measures; thus, moving to subjective outcome measures, we examine SVI. We find a main effect ($F(1, 144) = 113.64$, $p < .001$) of emotion on *relationship*, where people report better relationships with the *professional* agent ($M = 4.53$, $SD = 1.34$) than the *hostile* one ($M = 2.36$, $SD = 1.13$). We then find another main effect ($F(1, 146) = 7.02$, $p = .009$) of emotion on *self*, where people again report more positive feelings about themselves ($M = 5.80$, $SD = 0.84$) against the *professional* than *hostile* ($M = 5.40$, $SD = 1.03$) agent. For *Instrumental Outcome*, we find main effects of embodiment ($F(1, 146) = 6.09$, $p = .01$) and emotion ($F(1, 146) = 11.61$, $p < .001$); people report liking the outcome better ($M = 6.00$, $SD = 0.82$) against the *professional* agent than the *hostile* one ($M = 5.47$, $SD = 1.19$); and rate the outcome higher ($M = 5.91$, $SD = 0.90$) against the *disembodied* agent than the *embodied* agent ($M = 5.56$, $SD = 1.17$). Next, for *Process Fairness*, we see a main effect ($F(1, 145) = 92.81$, $p < .001$) for emotion, where people report the process as fairer against the *professional* agent ($M = 5.81$, $SD = 0.91$) than *hostile* ($M = 4.21$, $SD = 1.16$); we also see a main effect ($F(1, 145) = 3.99$, $p = .048$) of embodiment, where people against the disembodied agent ($M = 5.09$, $SD = 1.27$) viewed the process as more fair than those in the embodied condition ($M = 4.92$, $SD = 1.35$). Ultimately, we aggregate the entire SVI scale into a single value by averaging the four sub-scales and find another main effect ($F(1, 143) = 84, 68$, $p < .001$) of emotion, where again people report more positive scores against the *professional* ($M = 5.54$, $SD = 0.72$) agent than *hostile* ($M = 4.34$, $SD = 0.88$). Figure 6 shows the effects of *embodiment* and *emotion* on each of the SVI sub-scales[1].

5.2 Gender Effects

In this section, we examine the effect of gender on these measures further. We focus only on participants who identify as men or women, leaving us with 148 responses — 62% women and 38% men. Neither of the two removed were missing their transcript; thus, we again removed those without a transcript to analyze the dialog.

5.2.1 Language Sentiment. Now, considering gender, we again examine the language in the human-agent negotiation dialogs using the same technique in Section 5.1.1. As a manipulation check, we

[1]For this figure, as well as all others, we denote the significance scores of the adjusted p-value from Tukey's post-hoc test as $+p \leq .10$ $^{*}p \leq .05$, $^{**}p \leq .01$, and $^{***}p \leq .001$.

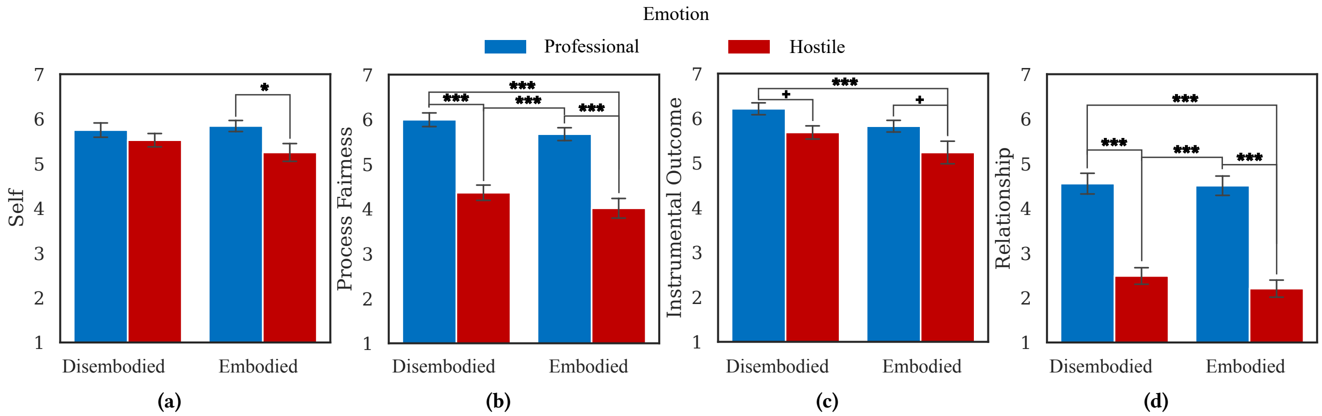


Figure 6: Overall subjective value index sub-scale scores by embodiment and agent emotion

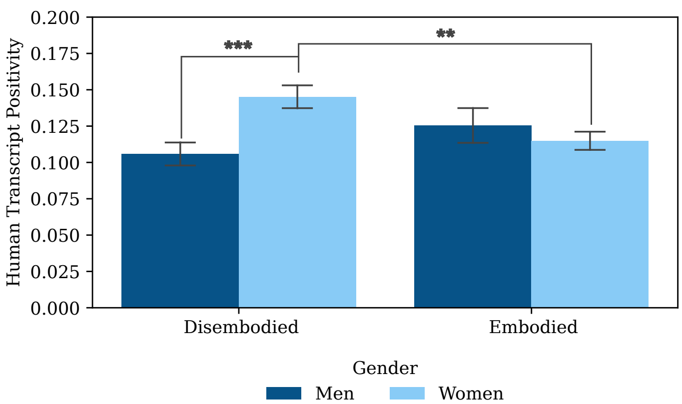


Figure 7: Human dialog positivity by embodiment and gender

see that the agent does not speak to men and women differently — there should not exist a significant difference in the sentiment of the agent's utterances between men and women. We perform a three-way analysis of variance on the agent's sentiment (emotion by embodiment by gender), which should have a main effect of emotion but no effect of gender — i.e., the agent should talk to men and women the same way. Indeed, we see a main effect of emotion ($F(1, 138) = 361.55$, $p < .001$), where the *professional* agent sends significantly more positive ($M = 0.22$, $SD = 0.04$) messages than the *hostile* ($M = 0.11$, $SD = 0.03$) one. Further, we see no effect of gender ($F(1, 138) = 0.83$, $p = .36$), verifying the agent does not differ significantly in its speech to men versus women.

Next, we analyze human utterances with the same three-way analysis of variance and find main effects of all three factors plus an interaction between gender and embodiment. Specifically, we see the *hostile* agent garnered significantly ($F(1, 138) = 16.76$, $p < .001$) lower ($M = 0.11$, $SD = 0.05$) sentiment responses from the human compared to the *professional* ($M = 0.14$, $SD = 0.05$) one; we see the *embodied* agent garnered significantly ($F(1, 138) = 5.17$, $p = .02$) lower ($M = 0.12$, $SD = 0.05$) sentiment than the *disembodied* one ($M = 0.13$, $SD = 0.05$); and women used significantly ($F(1, 138) = 8.22$, $p = .005$) higher ($M = 0.13$, $SD = 0.05$) sentiment messages than men ($M = 0.11$, $SD = 0.05$). Further, we see a significant ($F(1, 138) = 4.15$, $p = .04$) cross-interaction of gender and embodiment, where Tukey's posthoc test revealed that women against the disembodied agent send messages significantly higher in sentiment ($M = 0.15$, $SD = 0.05$) than women in the embodied condition ($M = 0.11$, $SD = 0.04$) and men in the disembodied condition ($M = 0.11$, $SD = 0.05$). Figure 7 illustrates this interaction.

5.2.2 *Tactics.* We now look at the *problem-solving*, *other-influence*, and *I-influence* in a three-way analysis of variance considering *embodiment*, *emotion*, and *gender*. We do not find a main effect of gender on *problem-solving*; however, we see a significant ($F(1, 140) = 4.94$, $p = .03$) cross interaction of gender and embodiment where men show less *problem-solving* in the disembodied condition ($M = -0.02$, $SD = 0.94$) compared to the embodied condition ($M = 0.38$, $SD = 0.83$), and women show more *problem-solving* in the disembodied condition ($M = 0.12$, $SD = 0.92$) compared to the embodied one ($M = -0.20$, $SD = 0.73$). We do not find gender effects for either *other-influence* or *I-influence*.

5.2.3 *Outcome.* While we see no effects on the objective measures (e.g., points), the subjective outcome metrics (SVI) show strong gender effects. We find a significant ($F(1, 138) = 4.09$, $p < .05$) interaction between *embodiment* and *gender* on *relationship*, where women report worse relationships with the *embodied* agent ($M = 3.14$, $SD = 1.62$) than the *disembodied* one ($M = 3.55$, $SD = 1.55$); conversely, men report better relationships with the *embodied* agent ($M = 4.22$, $SD = 1.67$) than the *disembodied* one ($M = 3.27$, $SD = 1.70$). We see another significant ($F(1, 139) = 5.72$, $p = .02$) interaction between *embodiment* and *gender* on *process fairness*, where women view the process as less fair against the *embodied* ($M = 4.66$, $SD = 1.39$) agent than the *disembodied* ($M = 5.18$, $SD = 1.25$) one; whereas men view the process as more fair against the *embodied* ($M = 5.54$, $SD = 1.08$) agent than the *disembodied* ($M = 5.00$, $SD = 1.33$) one. Lastly, we average the four sub-scales to analyze the overall SVI score where we see a third significant ($F(1, 137) = 4.84$, $p = .03$) interaction between *embodiment* and *gender* where, again, women perform worse against an *embodied*

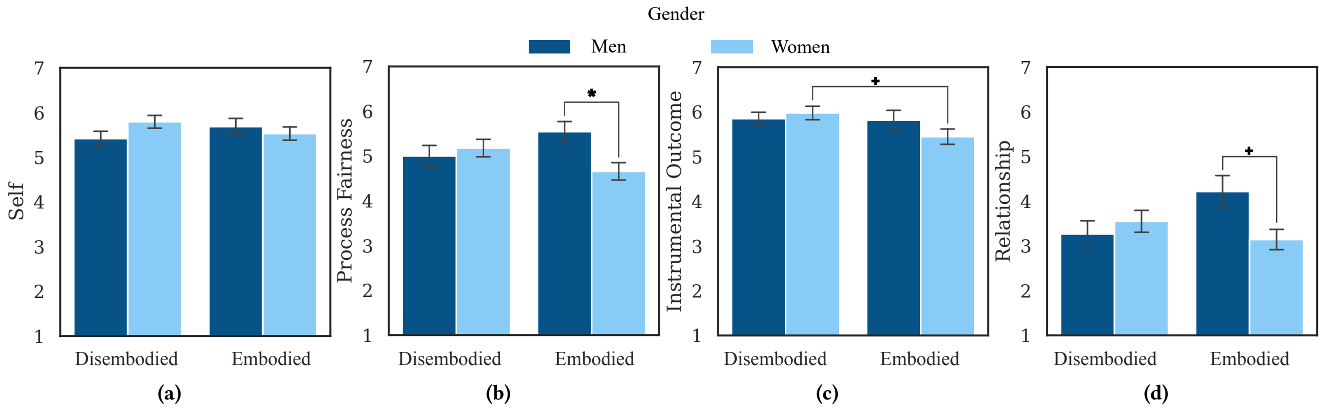


Figure 8: Overall subjective value index sub-scale scores by embodiment and gender

agent (M = 4.69, SD = 1.08) and better against a *disembodied* one (M = 5.12, SD = 0.94), while the reverse holds true for men — better in *embodiment* (M = 5.33, SD = 0.93) and worse in *disembodiment* (M = 4.89, SD = 0.92) embodiment. Figure 8 illustrates the interaction of *embodiment* and *gender* on the four SVI sub-scales. We find no significant gender effects for *self* nor *instrumental outcome.*

6 DISCUSSION

An initial analysis of the responses revealed several main effects and interactions of *embodiment* and *emotion.* We see that people use more positive language toward the *professional* agent compared to *hostile* and less positive to the *embodied* agent than the *disembodied* one. The agent's emotion affects various subjective measures such as *insight* and *problem-solving*, where participants perform better against the *professional* agent. We also find *both* embodiment and emotion effects on the SVI subscales, with higher scores for the *embodied* and *professional* agent — in line with past work [19].

Next, we incorporate *gender* into the analysis and find many instances where *gender* and *embodiment* interact in a cross pattern, such that if one gender group performs relatively better against an *embodied* agent compared to a *disembodied* one, the opposite will hold true for the other. For example, the participant's utterances rate comparatively higher in positive sentiment against the *embodied* agent if they identify as a man and comparatively lower if they identify as a woman, and this difference flips for the *disembodied* case. Also, men report relatively more *problem-solving* against the embodied agent, while women report relatively less. Using the subjective value index (SVI) we see this manifest when analyzing two of the sub-scales — *relationship*, and *process fairness.* However, these effects appear significantly only on subjective performance measures, while the actual score showed a trend. The agent's tendency to agree on a resolution may explain this, as all participants reach an agreement despite being able to walk away.

The masculine embodiment of the AI counterpart significantly disadvantaged women and benefited men *only* in terms of subjective measures; this may support work claiming women are more likely to hold an aversion to engaging in negotiation [4] since they perform approximately as well as men against the same AI — Johnson et al. [18] makes a similar claim that researchers should examine institutional externalities given the lack of objective performance differences against a controlled AI negotiator. Further, same as Johnson et al. [18], we find no significant effects in objective performance. Ultimately, these results highlight the need for caution in introducing embodiment to an experiment design. While our design disallows a conclusion on whether these gender differences stem from the embodiment itself or the gender of the embodiment, the fact that the agent appeared as a man implies the agent's gender promoted stereotypes in the interaction.

7 CONCLUSION & FUTURE WORK

Whereas the intelligent virtual agent community often assumes embodiment always benefits experiment design, we outline *why* and *where* this assumption may fall short. Specifically, men perform relatively better against an embodied agent, and women perform relatively worse across various metrics — e.g., textual analysis of the dialogs and subjective evaluations of the negotiation. As virtual human embodiment necessitates adopting a gendered and racial appearance, a researcher introducing this to their experiment may unintentionally allow social stereotypes to obscure their results.

The results presented in this paper emphasize the need to explore this question of embodiment in more depth in human-agent experiments. One limitation, as previously mentioned, is that we did not explicitly state to the participants whether they virtually negotiated against an AI- or human-driven counterpart, which may perturb the takeaways slightly. In future work, we will clarify this and measure to what extent a participant thought their counterpart was AI or human. Several directions exist for future work to unpack whether the AI's gender truly drives the effects outlined in this paper. First, we plan on soliciting crowd-sourced workers to rate the transcripts collected in this study while manipulating the agent's appearance (e.g., gender and race). Second, in a future run, participants will run through a version of this study in which we do not fix the agent's appearance as masculine but rather *manipulate* its gender. This will allow a deeper analysis of the presented interactions and allow a path to determine causation.

ACKNOWLEDGMENTS

This work is supported by the U.S. Government including the Air Force Office of Scientific Research, under grant FA9550-23-1-0320, and the Army Research Office under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Emily T Amanatullah and Catherine H Tinsley. 2013. Punishing female negotiators for asserting too much… or not enough: Exploring why advocacy moderates backlash against assertive female negotiators. *Organizational Behavior and Human Decision Processes* 120, 1 (2013), 110–122.

[3] Soroush Aslani, Jimena Ramirez-Marin, Jeanne Brett, Jingjing Yao, Zhaleh Semnani-Azad, Zhi-Xue Zhang, Catherine Tinsley, Laurie Weingart, and Wendi Adair. 2016. Dignity, face, and honor cultures: A study of negotiation strategy and outcomes in three cultures. *Journal of Organizational Behavior* 37, 8 (2016), 1178–1201.

[4] Linda Babcock and Sara Laschever. 2003. *Women don't ask: Negotiation and the gender divide.* Princeton University Press.

[5] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How Well Can LLMs Negotiate? NegotiationArena Platform and Analysis. *arXiv preprint arXiv:2402.05863* (2024).

[6] Tibor Bosse, Charlotte Gerritsen, and Jeroen de Man. 2016. An intelligent system for aggression de-escalation training. In *ECAI 2016*. IOS Press, 1805–1811.

[7] Justine Cassell. 2000. Embodied conversational interface agents. *Commun. ACM* 43, 4 (2000), 70–78.

[8] Jared R Curhan, Hillary Anger Elfenbein, and Noah Eisenkraft. 2010. The objective value of subjective value: A multi-round negotiation study. *Journal of Applied Social Psychology* 40, 3 (2010), 690–709.

[9] Jared R Curhan, Hillary Anger Elfenbein, and Heng Xu. 2006. What do people value when they negotiate? Mapping the domain of subjective value in negotiation. *Journal of personality and social psychology* 91, 3 (2006), 493.

[10] Andreea Danielescu. 2020. Eschewing gender stereotypes in voice assistants to promote inclusion. In *Proceedings of the 2nd conference on conversational user interfaces*. 1–3.

[11] Nabanita Datta Gupta, Anders Poulsen, and Marie Claire Villeval. 2013. Gender matching and competitiveness: Experimental evidence. *Economic Inquiry* 51, 1 (2013), 816–835.

[12] Nicole Davis, Nils Olsen, Vanessa G Perry, Marcus M Stewart, and Tiffany B White. 2023. I'm only human? The role of racial stereotypes, humanness, and satisfaction in transactions with anthropomorphic sales bots. *Journal of the Association for Consumer Research* 8, 1 (2023), 47–58.

[13] Lewis R Goldberg and Wayne F Velicer. 2006. Principles of exploratory factor analysis. *Differentiating normal and abnormal personality* 2 (2006), 209–337.

[14] Rosanna E Guadagno, Jim Blascovich, Jeremy N Bailenson, and Cade McCall. 2007. Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology* 10, 1 (2007), 1–22.

[15] Arno Hartholt, Kyle Mccullough, Ed Fast, Andrew Leeds, Sharon Mozgai, Tim Aris, Volkan Ustun, Andrew Gordon, and Chris Mcgroarty. 2020. Rapid Prototyping for Simulation and Training with the Rapid Integration & Development Environment (RIDE).

[16] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.

[17] Nusrath Jahan and Johnathan Mell. 2024. Unraveling the Tapestry of Deception and Personality: A Deep Dive into Multi-Issue Human-Agent Negotiation Dynamics. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 916–925.

[18] Emmanuel Johnson, Jonathan Gratch, Jill Boberg, David DeVault, Peter Kim, and Gale Lucas. 2021. Using Intelligent Agents to Examine Gender in Negotiations. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 90–97.

[19] Regina Jucks, Gesa A Linnemann, and Benjamin Brummernhenrich. 2018. Student evaluations of a (rude) spoken dialogue system insights from an experimental study. *Advances in Human-Computer Interaction* 2018, 1 (2018), 8406187.

[20] Kangsoo Kim, Luke Boelling, Steffen Haesler, Jeremy Bailenson, Gerd Bruder, and Greg F Welch. 2018. Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 105–114.

[21] Philipp Kulms, Nicole C Krämer, Jonathan Gratch, and Sin-Hwa Kang. 2011. It's in their eyes: A study on female and male virtual humans' gaze. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*. Springer, 80–92.

[22] Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale M Lucas, and Jonathan Gratch. 2024. Are LLMs Effective Negotiators? Systematic Evaluation of the Multifaceted Capabilities of LLMs in Negotiation Dialogues. *arXiv preprint arXiv:2402.13550* (2024).

[23] Sara Laschever and Linda Babcock. 2021. Women Don't ask: Negotiation and the gender divide. (2021).

[24] Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. 2006. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies* 64, 10 (2006), 962–973. https://doi.org/10.1016/j.ijhcs.2006.05.002

[25] Richard E Mayer and C Scott DaPra. 2012. An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied* 18, 3 (2012), 239.

[26] Jens Mazei, Joachim Hüffmeier, Philipp Alexander Freund, Alice F Stuhlmacher, Lena Bilke, and Guido Hertel. 2015. A meta-analysis on gender differences in negotiation outcomes and their moderators. *Psychological bulletin* 141, 1 (2015), 85.

[27] Jasmine McCormick and Wendy L Morris. 2015. The Effects of Stereotype Threat and Power on Women's and Men's Outcomes in Face-to-Face and E-mail Negotiations. *Psi Chi Journal of Psychological Research* 20, 3 (2015).

[28] Procheta Nag and Özge Nilay Yalçın. 2020. Gender Stereotypes in Virtual Agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Scotland, UK) *(IVA '20)*. Association for Computing Machinery, New York, NY, USA, Article 41, 8 pages. https://doi.org/10.1145/3383652.3423876

[29] Jeff Rickel and W Lewis Johnson. 1999. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied artificial intelligence* 13, 4-5 (1999), 343–382.

[30] Alice F Stuhlmacher and Eileen Linnabery. 2013. Gender and negotiation: A social role analysis. In *Handbook of research on negotiation*. Edward Elgar Publishing, 221–248.

[31] Alice F Stuhlmacher and Amy E Walters. 1999. Gender differences in negotiation outcome: A meta-analysis. *Personnel Psychology* 52, 3 (1999), 653–677.

[32] Henri Tajfel, Michael G Billig, Robert P Bundy, and Claude Flament. 1971. Social categorization and intergroup behaviour. *European journal of social psychology* 1, 2 (1971), 149–178.

[33] Larissa Z Tiedens and Alison R Fragale. 2003. Power moves: complementarity in dominant and submissive nonverbal behavior. *Journal of personality and social psychology* 84, 3 (2003), 558.

[34] Negin R Toosi, Zhaleh Semnani-Azad, Wen Shan, Shira Mor, and Emily T Amanatullah. 2020. How culture and race shape gender dynamics in negotiation. In *Research handbook on gender and negotiation*. Edward Elgar Publishing, 260–280.

[35] Joseph B Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research* 23, 1 (1996), 3–43.