Experiment Plan: Replicating the Halo Effect in AI-Generated Virtual Instructor

Group Members

Yadong Hou (NetID: yh2278)
Yifei Hu (NetID: yh2277)
Zhihao Mo (NetID: zm329)

Paper Citation

Full Citation: Nisbett, R. E., & Wilson, T. D. (1977). The Halo Effect: Evidence for Unconscious Alteration of Judgments. *Journal of Personality and Social Psychology, 35*(4), 250-256. https://doi.org/10.1037/0022-3514.35.4.250

Paper Link: https://psycnet.apa.org/doiLanding?doi=10.1037/0022-3514.35.4.250

Rationale for Paper Selection

- The halo effect remains highly relevant in today's interactions, particularly as AI-generated virtual agents become increasingly prevalent in education, customer service, and digital interfaces.
- Extending this classic study to AI-generated instructors addresses a question: do humans exhibit the same unconscious biases toward virtual agents as they do toward real humans.

Original Experiment Summary

Hypothesis: study tested two hypotheses:

1. global evaluations of a person can alter evaluations of that person's specific attributes, even when sufficient information exists for independent assessment.
2. individuals are unaware of this influence, potentially misattributing the direction of causality.

Sampling & Recruitment: Participants were 118 students (62 males, 56 females) enrolled in psychology courses at the University of Michigan. They participated in groups of 6-17 students.

Sample Size: N =118

Treatments/Conditions: Participants viewed one of two video interviews with the same college instructor, a native French-speaking Belgian with a European accent.

In the warm condition, the instructor appeared friendly, respectful of students, flexible in teaching approach, and enthusiastic.

In the cold condition, the same instructor appeared distant, distrustful toward students, rigid, and doctrinaire.

All participants first watched a neutral filler interview before seeing the target instructor.

Independent Variable: Instructor demeanor (warm/friendly vs. cold/distant)

Dependent Variables:

- Overall likability of the instructor
- Ratings of physical appearance
- Ratings of mannerisms
- Ratings of accent

Measurement: Likability was measure on an 8-point scale from "like extremely" to "dislike extremely." Specific attributes were rated on an 8-point scale from "extremely appealing" to "extremely irritating." Additionally, participants reported whether they believed their global evaluation influenced their attribute ratings.

Key Findings: Participants in the warm condition rated the instructor's appearance, mannerisms, and accent significantly more appealing than those in the cold condition, despite these attributes being objectively identical.

Participants were unaware of this influence. Those in the cold condition actually believed the opposite causality, that dislike of specific attributes caused their negative global evaluation.

Technological Adaptation

We will adapt this experiment to examine whether the halo effect occurs with AI-generated virtual instructors, let this study relevant to AI educational technologies and human-AI interaction.

Implementation: We will use AI video generation platforms (e.g., Sora, Veo3) to create two videos featuring the same AI-generated virtual instructor. The virtual instructor will deliver identical content about a psychology topic but with contrasting demeanors:

- **Warm condition:** The AI instructor will display friendly facial expressions, warm vocal tone, welcoming gestures, and positive body language
- **Cold condition:** The same AI instructor will exhibit neutral/negative facial expressions, monotone voice, minimal gestures, and distant body language

Variables:

- **Independent Variable:** AI instructor's demeanor (warm vs. cold)
- **Dependent Variables:**
    - Overall likability ratings
    - Perceived attractiveness of the avatar's visual design
    - Perceived appeal of the avatar's mannerisms/animations
    - Perceived pleasantness of the voice/vocal tone

Measurement: Participants will complete the same 8-point rating scales used in the original study, adapted for virtual agents. We will also assess awareness by asking whether global impressions influenced specific ratings.

Sample: We aim to recruit 20-30 participants, randomly assigning them to warm or cold conditions.

Significance: This adaptation addresses whether humans exhibit unconscious bias toward AI-generated agents.

Expected Results: Mock Data Figures

Based on the original study's findings, we anticipate that participants who interact with the warm AI instructor will rate all specific attributes more positively than those in the cold condition.

Figure 1: Effect of AI Instructor Demeanor on Attribute Ratings

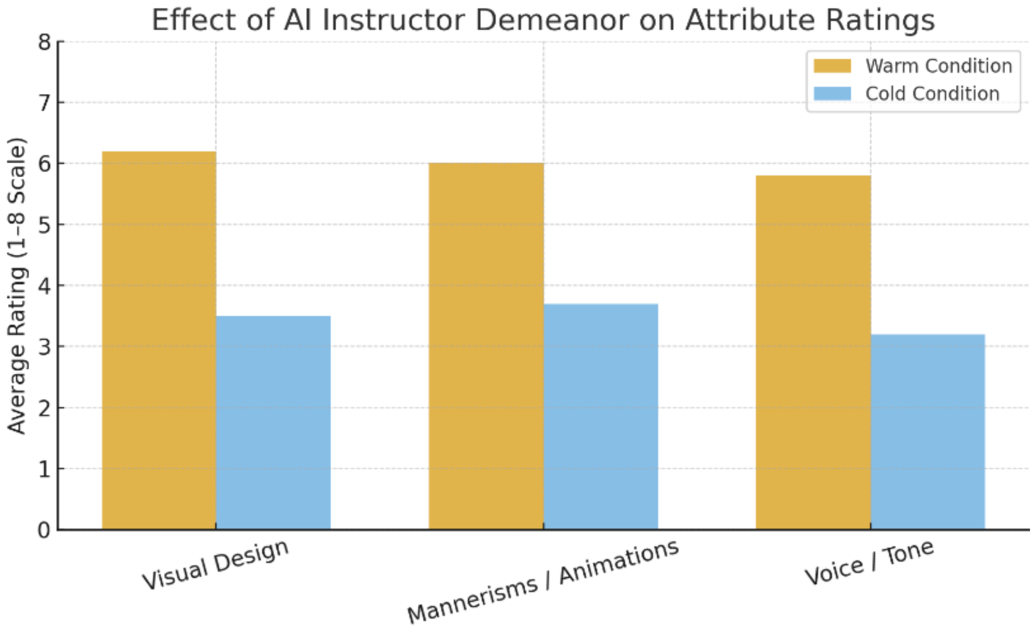

Figure 1 shows that people clearly liked the warm AI instructor a lot more than the cold one. Even though both videos showed the exact same avatar and content, participants gave the warm version much higher ratings for how it looked, moved, and sounded. On average, the warm instructor scored around six out of eight on all measures, while the cold one only got around three to four. In short, when the AI seemed friendlier and more enthusiastic, people thought *everything* about it was better—even things that shouldn't have changed.

Figure 2: Percentage of Participants Rating Attributes as "Appealing" vs. "Irritating"

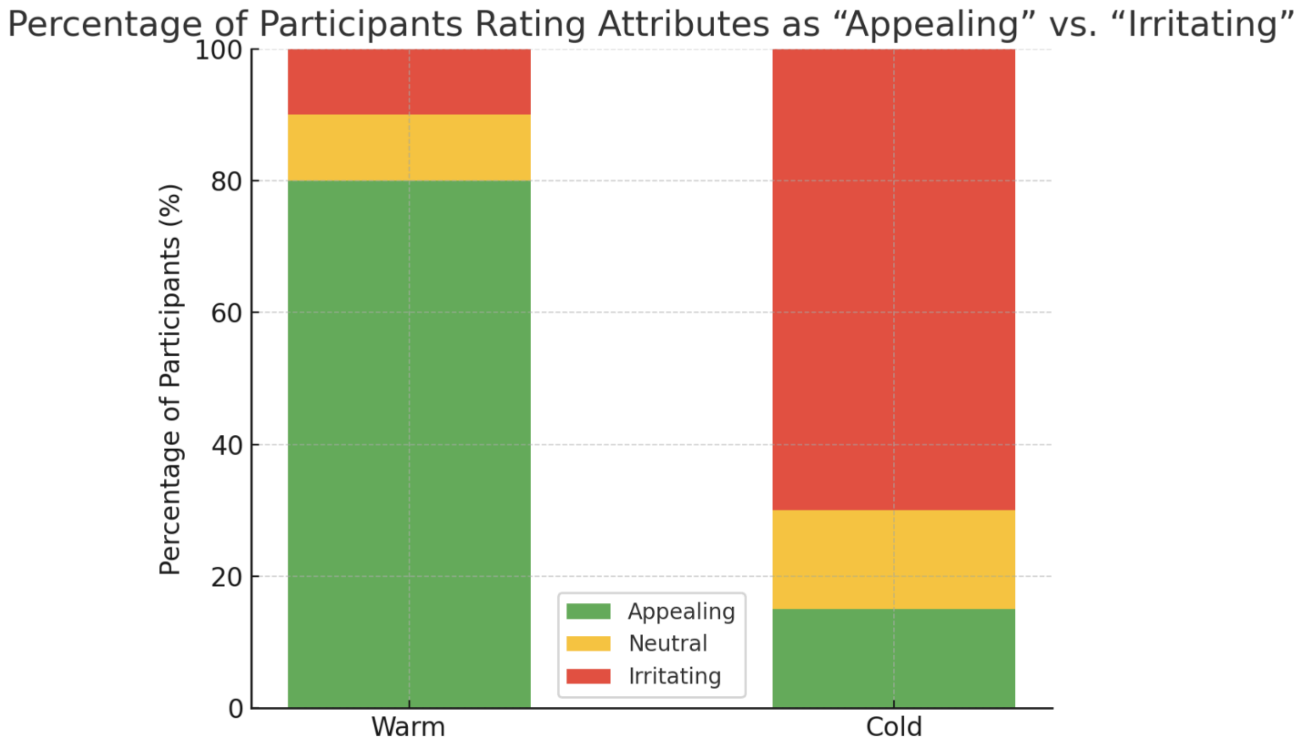

Figure 2 tells a similar story. Most people who saw the warm instructor thought its design and voice were appealing—about 80 percent of them—while almost everyone who saw the cold instructor found it irritating. The difference is huge and really highlights the halo effect: if we like the overall vibe of an AI, we tend to assume all its traits are good too; if we don’t, we judge everything more harshly.