International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

The Role of Perceived Judgement in Self-Disclosure: A Study of AI and Human Interaction

Amulya Thomas

Abstract

The study examines the effect of interaction medium (AI/human) on the levels of self-disclosure. The study is grounded in theories like the Online Disinhibition Effect, Theory of Mind and Social Desirability Bias, which suggest that the context in which communication occurs influences openness. A sample of 60 participants was divided into 2 conditions: AI and human interaction respectively. Responses to a structured questionnaire were coded on a 0-3 self-disclosure scale. It was analysed using the chi-square test. The results revealed a statistically significant association between the interaction type and level of disclosure, $\chi^2(3) = 9.14$, $p < 0.05$. Findings suggest AI-based interactions result in both higher and lower levels of self-disclosure while human interactions produce more moderate, generic responses. The study highlights the role of perceived safety and lack of judgement in shaping disclosure behavior.

Keywords: Self-disclosure, Artificial Intelligence, Online Disinhibition Effect, Social Desirability Bias, Theory of Mind, Communication Psychology

Introduction

Self-disclosure is the revealing of personal information such as thoughts, feelings, and opinions (Richard L Archer; Joseph A Burleson et al., 1980). It plays a significant role in developing interpersonal relationships and improving one's sense of emotional wellbeing. Humans disclose information based on factors like trust, nature of the relationship and perceived judgement. Individuals control how much personal information they share with others consciously or unconsciously. This is done to maintain their social image, protect their privacy or avoid criticism.

The concept of social desirability bias, also known as SDB, explains this phenomenon. In SDB, people modify their responses to be viewed more desirably by others. This includes disclosing only surface level information about oneself, or omitting information of emotional charge.

However, with rapid advancements in technology, new modes of interaction have emerged, particularly with Artificial Intelligence (AI) systems. Unlike human interaction, AI systems are often perceived as non-judgemental, anonymous and accepting, which may reduce the fear of judgement while disclosing information. The fear of judgement may stem from how we think others will react. We believe others will judge or stereotype us, or that they won't understand us. The Theory of Mind (ToM) explains this. It refers to our ability to attribute mental states such as beliefs, desires, intentions and emotions to others.

The shift in interaction after the rise of technology can also be understood through the Online Disinhibition Effect (Suler et al., 2004). It states that individuals tend to express themselves more in digital environments due to reduced social constraints. Factors like privacy, immediate response and perceived safety can encourage them to open up more about sensitive information.

Studies have found that people may feel less judged while disclosing to a CUI compared to a human (Rune Moberg Jacobsen, Samuel Rhys Cox, Carla F Griggio, Niels van Berkel, et al., 2025). People may feel more judged with higher levels of social presence. Pickard et al., found that people had a higher chance of disclosing shame to a voice agent that was disembodied as compared to an embodied CA or even a human interviewer.

Acknowledgement recognizes and affirms the other person's lived experiences. Applied to an AI context, studies of empathic concern often use acknowledgement to validate individuals who communicate personal information to chatbots. Phrases like "I understand how frustrating this must be to you" increase the user's perception of how much the chatbot supports and understands them (Liu and Sundar, et al., 2018).

At the same time, it should be noted that AI interactions need not always be deeper or honest. Some may feel skeptical or wary while interacting with AI systems.

There are critical differences between human empathy and artificial empathy as seen in AI or CUI models. AI agents cannot feel or experience like humans do. They can only stimulate human empathy by displaying pseudo-mental features of empathy (Airenti, et al., 2015).

This makes artificial empathy possible through computational algorithms. Emotional contagion, while natural to humans, has to be programmed into an AI system through learning accumulated data. While not fully aware of the logistics, some individuals continue to be careful of disclosing personal information to AI systems precisely due to the lack of true understanding and human connection.

Hence, the impact of such interactions is not absolute, but varies depending on the individual and context. This study aims to differentiate and understand how self-disclosure differs when individuals respond in two conditions: interaction with an assumed AI chatbot, and interaction with an assumed human. By analyzing levels of disclosure among individuals in each question in the two scenarios, the study seeks to find out whether people disclose more sensitive information to AI or whether traditional forms of communication are preferred.

Aim of the Study

To examine the effect of the mode of interaction (AI vs human) on levels of disclosure. The study specifically seeks to compare how individuals vary in depth and nature of personal information while responding to questions in an AI-mediated versus a perceived human context.

Hypothesis

Null Hypothesis (H_0)

There is no significant relationship between the type of interaction (AI vs human) and the level of disclosure.

Alternative Hypothesis (H_1)

There is a significant relationship between the type of interaction (AI vs human) and the level of disclosure.

Variables:

Independent Variable (IV)

Operational Definition:

The condition under which participants responded to the survey:

- AI condition: Participants believed they were responding to an AI chatbot.

- Human condition: Participants believed they were responding to a human.

Dependent Variable (DV)

Level of disclosure

Operational Definition:

The extent to which participants revealed personal information in their responses, which was measured using a 0-3 disclosure coding scale where:

- 0 = No disclosure
  Eg: "idk", "no", no responses, refusal to answer
- 1 = Low disclosure
  Surface level, generic, factual responses with little to no personal detail
- 2 = Moderate disclosure
  Personal responses with some emotional expression
- 3 = High disclosure
  Disclosing deeply personal information, significant life experiences, showing emotions

Controlled Variables

- Same set of questions was provided for both conditions
- Same number of participants in both groups
- Same response format (open-ended survey)
- Same coding criteria used for both conditions

Methodology

Sample

The study was conducted on a sample of 60 participants. Participants were divided into two groups:

- 30 participants in the AI condition
- 30 participants in the human condition

The sample included participants from a diverse age group ranging from 13 to above 30 years of age.

Tools Used

- A self-constructed questionnaire that consists of 14 open-ended questions related to personal experiences, emotions and self perception.
- A 0-3 self disclosure coding scale
- Statistical analysis using the chi-square test

Procedure

Participants were asked to respond to a set of open-ended questions under one of the two conditions:

- In the AI condition, participants believed they were interacting with an AI chatbot
- In the human condition, participants believed responses would be read by a human.

All responses were collected and then coded based on the level of disclosure using a standardized 0-3 scale.

The frequency of each disclosure was calculated for both conditions. A chi-square test of independence was then conducted to determine whether there was a significant relationship between interaction type and level of disclosure.

Results

The chi-square test of independence was conducted to examine the relationship between interaction medium and level of self-disclosure. The results indicated a statistically significant association, $\chi^2(3) = 9.14$, $p < 0.05$, suggesting that the type of interaction influences the depth of personal information shared by individuals.

The distribution of responses showed clear differences between conditions. In the AI interaction group, responses showed relatively higher frequencies in both low (level 1) and high (level 3) disclosure categories. In contrast, the human/stranger group demonstrated a greater concentration of moderate disclosure (level 2).

Discussion

The findings suggest that interaction medium plays a meaningful role in shaping self-disclosure behaviour. The higher occurrence of extreme responses (low and high disclosure) in the AI condition may be explained by the Online Disinhibition Effect (Suler, et al., 2004), which proposes that individuals feel less constrained and less judged in digital environments, thereby increasing both openness and detachment.

AI interactions may also be perceived as emotionally neutral and non-judgmental, which can encourage some individuals to share deeply personal experiences. However, the presence of low disclosure in the same condition indicates that AI does not uniformly increase openness; instead, it may also lead to minimal engagement in some individuals due to perceived lack of human connection.

In contrast, the human/stranger condition showed a dominance of moderate disclosure, reflecting the influence of social norms and social desirability bias. Individuals tend to regulate their responses in human interactions to maintain a socially acceptable image and avoid negative evaluation. This results in more balanced and controlled self-disclosure.

Overall, the findings indicate that while AI can create a psychologically safer space for some individuals to express themselves more freely, human interaction still plays a strong regulatory role in maintaining socially moderated communication patterns.

Conclusion

The present study concludes that there is a significant relationship between interaction medium and level of self-disclosure. Participants displayed different patterns of disclosure depending on whether they believed they were interacting with an AI system or a human/stranger. AI interactions led to more extreme forms of disclosure, while human interactions encouraged more moderate and socially regulated responses.

However, the effect was moderate in strength, indicating that while interaction context influences self-disclosure, it does not fully determine it. Individual differences and question content also play an important role in shaping disclosure behaviour.

Overall, the study highlights the evolving nature of communication in digital environments and suggests that AI has the potential to reshape how individuals express personal thoughts and emotions, though traditional social influences remain significant.

Bibliography

1. Richard L Archer; Joseph A Burleson et al., 1980. The Effect of Timing of Self Disclosure on Attraction and Reciprocity
2. Rune Moberg Jacobsen, Samuel Rhys Cox, Carla F Griggio, Niels van Berkel, et al., 2025. Chatbots for Data Collection in Surveys: A comparison of four theory-based interview probes. In proceedings of the 2025 CHI conference on human factors in computing systems
3. Liu and Sundar et al., 2018. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot
4. Airenti et al., 2015. The Cognitive Bases of Anthropomorphism