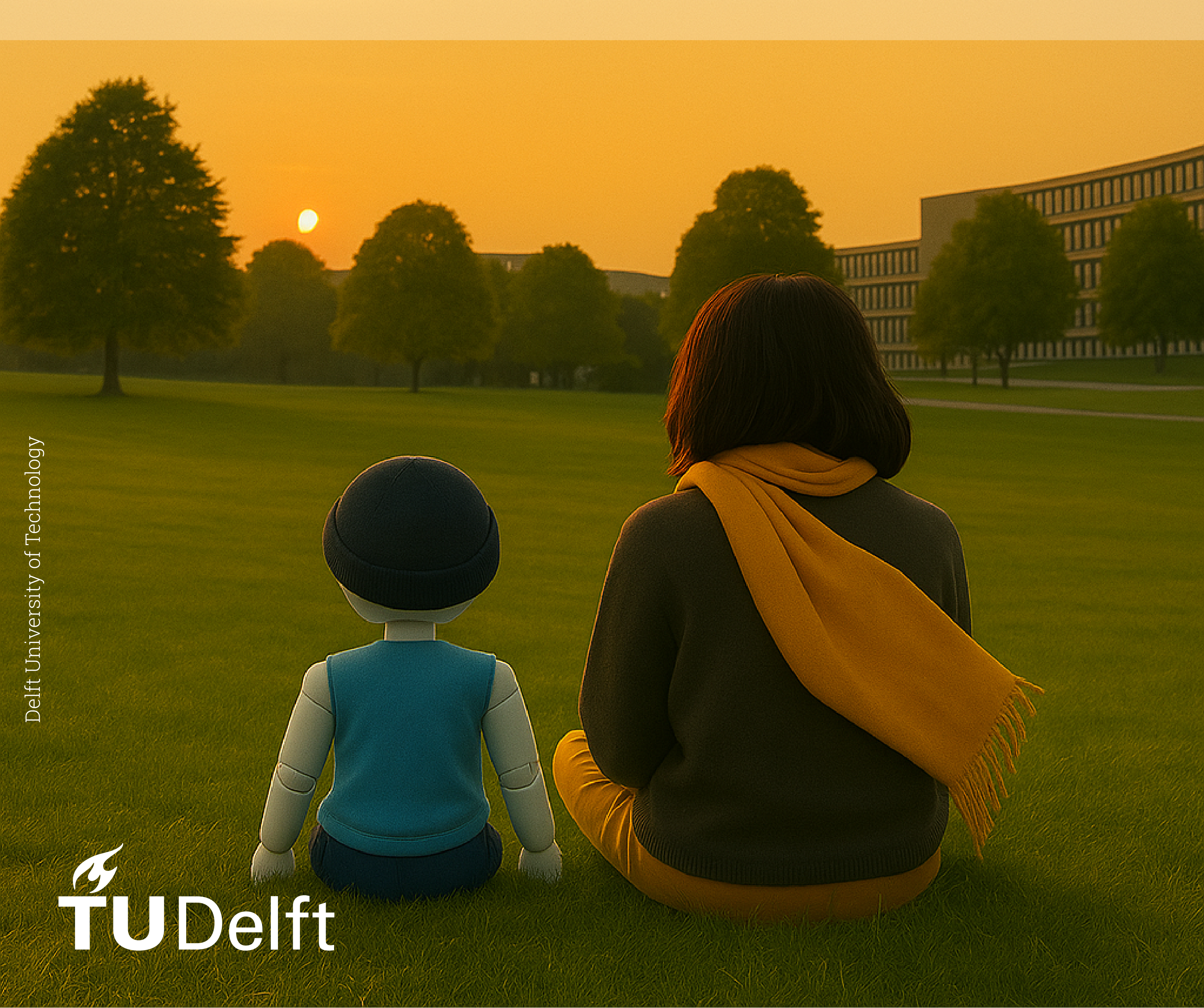
Understanding The Development of Human Trust in Social Robots

MSc Computer Science Thesis

Chih-Wei (Charlotte) Ning

Delft University of Technology



Understanding The Development of Human Trust in Social Robots

by

Chih-Wei (Charlotte) Ning

to obtain the degree of MSc Computer Science

at the Delft University of Technology,

to be defended publicly on Tuesday July 8, 2025 at 13:00.

Student number:	5914558	
Project duration:	September 5, 2024 – July 8, 2025	
Thesis committee:	Dr. Myrthe L. Tielman	(supervisor)
	Prof. Dr. Mark A. Neerincx	
	Dr. Bernd J. W. Dudzik	(external)
Daily supervisor:	Carolina Centeio Jorge	

The work in this thesis was conducted at the:

Interactive Intelligence Group
Department of Intelligent Systems
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Abstract

As robots and virtual agents are increasingly envisioned as long-term companions rather than simply tools, it becomes essential to ensure that human–robot relationships are grounded in appropriate forms of trust. This study investigates how cognitive and affective dimensions of trust develop differently over time in social human–robot interaction. We conducted a 2 (social attitude: social, baseline) × 3 (time: t_1, t_2, t_3) mixed-design user study using a novel, card-based conversational task designed to encourage trust formation. Results show that while cognitive trust remained stable over time, affective trust increased gradually across repeated interactions. Moreover, social cues enhance both cognitive and affective trust. These findings provide empirical support for the theoretical distinction between cognitive and affective trust, offering new evidence that affective trust develops more slowly, consistent with interpersonal trust theories.

Acknowledgment

It took me approximately one year to come up with and complete this thesis project, but it took 28 years for the universe to bring me to this exact place and moment, where I am able to share my work with all of you. This journey has only been possible thanks to the support of so many people.

I'd like to start this long list of gratitude with my supervisory team. To my advisor, Dr. Myrthe Tielman: thank you for always encouraging me to play around every random idea I have. I still remember our first encounter, when there were only four people in the lecture hall in a cold winter morning. I'm so glad I was one of them, even though I'm never an early bird. My daily supervisor Carolina Centeio Jorge, thank you for always paying detailed attention to my work and giving me valuable feedback. To Ruben Verhagen and my labmates, Harmen and Stelios: thank you for being part of the team, our weekly discussions have always been inspiring. I'm also grateful to Prof. Dr. Mark Neerincx and Dr. Bernd Dudzik for being on my committee, and to the TU Delft Justus & Louise van Effen Excellence Scholarships for making this two-year journey possible in the first place.

During this project, I was lucky to receive generous support from many professionals. Thank you, Shambhawi Pal and Ruud de Jong, for introducing me to Navel, the cutest social robot. Thank you, Mohammed Al Owayyed and Dr. Franziska Burger, for answering my questions on intimacy ratings. To Helen, Patrice, and Yolan: thank you for being my Tarot consultants.

To my family, my mother Gem, my sister Charlie, and my sister in-law Katherine: thank you for being my safe fortress, giving me the courage to take risks, knowing you'll always have my back. I dedicate this thesis to my father Ning, whom I've always admired as a researcher. It's been almost a year since you left us. Without physical constraints now, I believe you'll be there at my defense, and I know you'd be proud.

To my dearest friends Teng, Kang, and Peng, thank you for being both my shoulders to cry on and part of my sweetest memories in Delft. My experiment can never run so smoothly without Pei-Yu's laptop and Teng's green Kuai Kuai — thank you for that. To all the friends who helped with my study or simply supported me, even if your name isn't explicitly written here, your presence has meant a lot to me. My lifelines across the world: Rilak, Jason, Hummingbird, Nkhg, Handsome, Kaitlyn, Vincent, and YiChi, thank you for always picking up the phone; with you, I am never alone. Thank you, my therapist Mason, for helping me stay grounded through uncertain times.

I want to thank myself for working so hard in the past two years, not just on the research but also on carrying myself through all the bittersweetness. I'm grateful for this thesis project itself, for reminding me that I am valuable and autonomous. I am grateful for every piece of sunlight and every blooming flower on this land, for embracing me as an international. And thank you to the people of Taiwan, for guarding my homeland while I'm away.

Finally, this is a project about relationships and companionship. Anyone who has ever walked by my side, whether briefly or for a long time, has shaped who I am today, and inspired how I think about what relationships mean to individuals and to the society. One day, someone might include a social agent in their thesis acknowledgments. What would that mean? And what should we do about that? I don't know yet, but I'm thrilled to find out.

Chih-Wei (Charlotte) Ning
Delft, June 2025

Contents

List of Figures

List of Tables

1 Introduction

1.1. Motivation

1.1.1. Long-term Relationships with Social Robots

There is a paradigm shift in social robotics research, where robots are increasingly expected to take on roles as social actors rather than merely functional tools [60]. Social robotics has emerged from multiple research fields as a response to the shortage of manpower in caregiving, teaching, service industries... etc [6, 23, 60, 35]. Some robots assist the elderly with simple daily chores and engage them in artistic or intellectual activities [23], while others support students in educational games [35]. These robots are carefully designed and monitored to not only accomplish specific tasks, but also to provide a sense of companionship [60]. The social support provided by robots has been reported to reduce users' stress, anxiety, and loneliness [48, 22].

Robots' role as social companions has been further emphasized with recent advancements in AI technology, particularly through personalization and large language models. Data-driven approaches now allow an agent to be highly tailored to individual user and to evolve from a one-time servant to a lasting companion. The maturation of large language further enables conversational agents to create more organic conversations. These advancements have moved far beyond the research lab and are now commonplace in products like OpenAI's GPT and Amazon's Alexa.

With these increasingly natural interactions, forming long-term relationships with artificial agents has become more feasible. For many, talking to a virtual agent has become as natural as texting a friend. Narratives once confined to science fiction, such as those depicted in the film *Her*, have begun to materialize in real life. People now turn to social agents not only for advice, but also for emotional companionship and support [45].

1.1.2. Appropriate Trust in Human-Robot Relationships

Trust plays an essential role in both interpersonal and human–robot relationships [33, 62]. It is commonly defined as an attitude that the trustor holds towards the trustee, when the trustor feels positive about relying on the trustee, despite being in a situation characterized by uncertainty and vulnerability [19, 31]. In this context, trusting someone means willingly placing oneself in a potentially risky or dependent position. Researchers in HRI (human-robot interaction) have thus focused on fostering appropriate trust between humans and artificial agents [44]. That is, a level of trust that avoids both overtrusting (and misuse a system) and undertrusting (and fail to benefit from it).

Various models (e.g. [41, 40]) conceptualized trust as a multidimensional construct, often distinguishing between ability-based and relational components. One example is the cognitive–affective model proposed by McAllister [42]. While relational dimensions have long been recognized as essential for forming close interpersonal bonds [33, 62, 39], they remain relatively underexplored in HRI. Most studies still prioritize ability-related factors such as competence or reliability [44]. This focus was understandable when artificial agents were primarily framed as tools, but became insufficient as the paradigm shifted

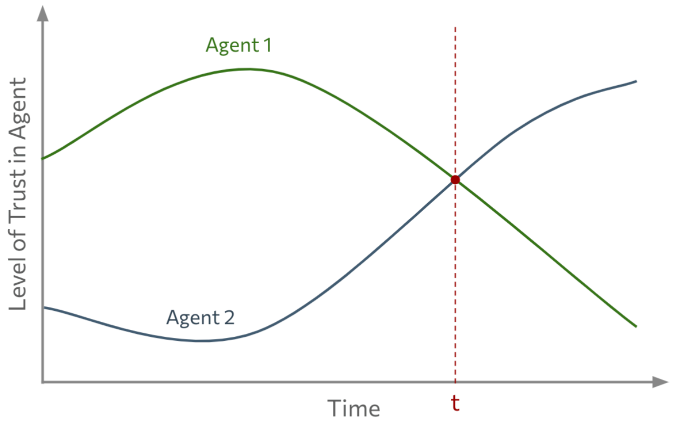


Figure 1.1: Human's trust in agents develop dynamically. For example, the level of trust in both agents are identical at time *t*. However, the dynamics are completely different. (Figure adapted from Guo et al. [25])

toward socially engaging systems.

However, in emotionally meaningful, long-term HRI scenarios (see Section 1.1.1), the risks of inappropriate affective trust become especially salient. Excessive trust can lead to overreliance, emotional dependency, or even situations that should be better addressed by professional care. These concerns underscore the need to better understand and support appropriate forms of affective trust.

1.1.3. The Development of Multidimensional Trust

Ensuring appropriate trust in HRI requires attention to its *temporal* nature, as trust development is an ongoing calibration process built through users' accumulated experiences [29, 54]. As illustrated in Figure 1.1, incorporating temporal information offers a more comprehensive understanding of this process. Moreover, when emotionally meaningful relationships are the focus, time becomes inherently crucial: deeper emotional bonds take time to form and strengthen [33, 32, 31, 55]. Therefore, trust should be viewed as a dynamic process rather than a fixed measurement taken at a single moment in time [25, 44, 54].

Prior research suggests that distinct trust dimensions may follow distinct cognitive pathways [31, 46, 29] and thus form non-identical developmental trajectories [55, 62]. Affective trust is often assumed as gradually emerging through repeated engagement, while cognitive trust tends to form earlier through rational assessments of competence [55, 33]. This premise comes from interpersonal literature and is widely adopted in HRI studies [62, 16]. However, direct empirical evidence in the context of socially engaging HRI remains limited.

This research, therefore, aims to investigate how different dimensions of trust may evolve differently over time, and whether a similar transition from cognitive-based to affective-based trust occurs in HRI as it does in interpersonal relationships. We hope the findings will help bridge this gap and clarify inconsistencies in the literature. For instance, two studies with similar designs reached opposite conclusions on whether competence or benevolence has stronger influence on trust in social robots [6, 23]. While the former involved around 15 minutes of interaction and the latter only a brief exchange, we argue that *time* might be the key factor. If affective responses require more time to emerge, this discrepancy could plausibly explain the divergence in findings.

In summary, while social robots may offer supportive experiences and even foster interpersonal-style relationships, such relationships must be grounded in calibrated, appropriate trust. We must understand how different dimensions of trust may follow distinct developmental trajectories over time in human–robot interaction. This research aims to contribute to that understanding, advancing toward a more ethically informed and emotionally safe hybrid society.

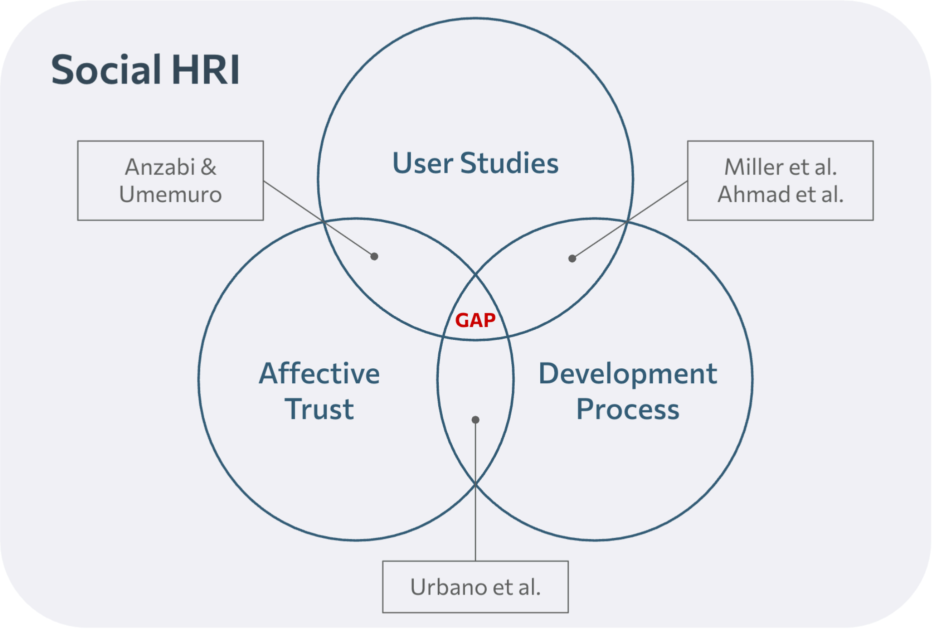


Figure 1.2: Literature exists at the intersection of each pair among affective trust, trust development, and experimental HRI with users. This study targets the underexplored overlap of all three.

1.2. Research Goal

The focus of this research targets the interceptions of three aspects, as illustrated in Figure 1.2:

1. **Affective Trust**: Trust is treated as a multidimensional construct, with particular emphasis on the affective dimension in McAllister's cognitive–affective model [42].
2. **Development Process**: Trust is regarded as a dynamic process rather than a static variable. This research seeks to examine the trajectory of trust as it unfolds over time.
3. **User Studies**: Trust development is investigated through controlled lab experiments in which users engage with social robots over multiple interactions.

In the domain of social HRI, prior studies have explored each pairwise combination of the three focuses. Urbano et al. have modeled benevolence as a separate dimension of trust in their dynamically-evolved computational models [62]. Anzabi and Umemuro have applied the cognitive–affective trust model in experimental studies on social HRI [6, 5]. As for development process, multiple scholars have investigated trust dynamics in repeated measured user studies [46, 1]. However, the intersection of all three aspects remains underexplored, leaving a gap that this research aims to address.

1.2.1. Research Questions and Hypotheses

The main research question that guides this thesis is:

- **RQ:** How does multidimensional trust in a social robot develop over time?

As discussed in Section 1.1.3, this research is grounded in the assumption that while cognitive trust tends to form early in an interaction, affective trust typically takes more time to develop. Based on this, we propose the first sub-question:

- **RQ(a):** What are the developmental trajectories of cognitive and affective trust? How do they differ over time?

This question is exploratory in nature. Since trust is commonly understood as a dynamic, continuously calibrated process shaped by interaction experiences [29], we do not make directional assumptions about the specific shape of either trajectory. However, we hypothesize temporal difference between dimensions:

- **H(a):** Affective trust takes more time to develop than cognitive trust.

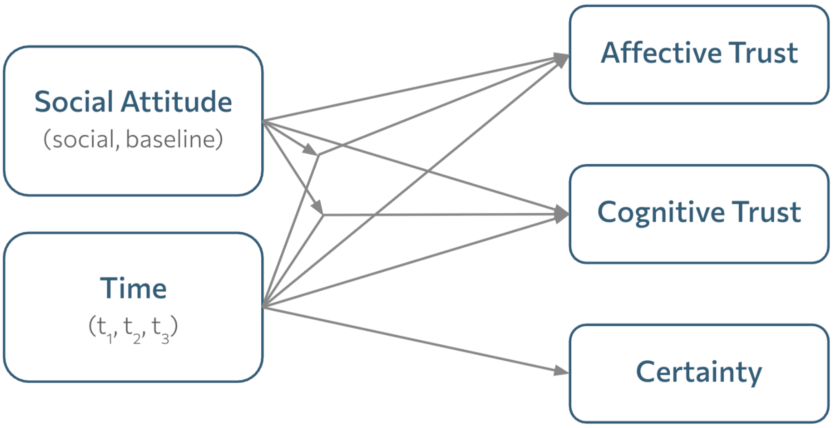


Figure 1.3: The conceptual model of the variables in this research. The left column shows the independent variables: *social attitude* (between-subject) and *time* (within-subject). The right column shows dependent variable: *affective trust*, *cognitive trust*, and *certainty*. The arrows depict the hypothesized effect relationships. The forked arrows that link both *social attitude* and *time* to *affective/cognitive trust* stand for interaction effects

We will answer the research questions through a controlled study, which introduces *social attitude* as a second independent variable alongside time. For a robot to function as a supportive social actor, it must convey appropriate social signals, enable empathetic interactions, and potentially foster emotional connection and trust. By manipulating the robot's social attitude, we aim to replicate prior findings suggesting that social behaviors from a robot increase trust [23, 50, 6]. More importantly, we treat the non-socialized robot as a baseline to understand how trust unfolds in the presence of social cues. This leads to the next research question:

- **RQ(b):** How is the development of cognitive and affective trust affected by a robot's social attitude?

We expect a more socially expressive robot to elicit higher levels of both cognitive and affective trust, and to interact positively with time to facilitate greater growth in trust. Additionally, following the assumption that affective trust emerges more slowly than cognitive trust, we expect the timing of group differences to differ between the two trust types. We thus propose:

- **H(b.1):** Social attitude has a positive effect on the cognitive trust in a social robot.
- **H(b.2):** Social attitude has a positive effect on the affective trust in a social robot.
- **H(b.3):** The positive effect of social attitude on affective trust emerges at a later stage comparing to the effect on cognitive trust.

Drawing from the three-layer model of human-automation trust [26], users may enter the interaction with pre-existing trust attitudes which they may feel uncertain about. However, this study focuses on dynamic-learned trust, which develops through direct interaction. To help disentangle trust that evolves from interaction versus trust that is carried in, we additionally measure participants' certainty in their trust assessments over time. This leads to the following research question:

- **RQ(c):** How does users' certainty in their trust assessments evolve over time?

Dynamic learned trust is continually calibrated as users compare observed system behavior with the initial expectations [29]. Throughout repeated interactions with the robot, users naturally accumulate more information that assists trust evaluations. This leads to the corresponding hypothesis:

- **H(c.1):** Users are more certain about their judgment in cognitive trust over time.
- **H(c.2):** Users are more certain about their judgment in affective trust over time.

Although we do not propose specific hypotheses about the impact of social attitude on certainty, we will still explore potential group differences.

Figure 1.3 shows the hypothesized relationship among all relevant variables.

1.2.2. Contributions

The contributions of this research include:

1. An investigation and comparison of the development of cognitive and affective trust as distinct dimensions in social HRI.
2. Providing empirical support that cognitive trust emerges earlier, while affective trust takes more time to develop.
3. Designing and implementing a novel interaction task and stimuli, which focus on emotional meaningful relationships and effectively elicit both cognitive and affective trust.
4. Bridging gaps between affective trust dimension, temporal trust development, and experimental methodology in social HRI.

1.2.3. Report Structure

This chapter has outlined the motivations and research goals. The remainder of this thesis is structured as follows: Chapter 2 provides a comprehensive review of relevant literature. Chapter 3 introduces a novel experimental paradigm: a card-based divination task, and describes its design and implementation. This paradigm is employed in the empirical study presented in Chapter 4, which aims to answer the research questions. The study results are reported in Chapter 5 and further discussed in Chapter 6. Finally, Chapter 7 concludes the thesis and outlines potential directions for future work.

2 Background

As discussed in Section 1.2, this thesis explores the intersection of affective trust, dynamic trust development, and experimental HRI. This chapter presents a comprehensive review of relevant literature, structured around these three focal areas.

The first part (Section 2.1 and Section 2.2) outlines the theoretical foundations of trust, drawing from research in interpersonal relationships, human–machine interaction, and social robotics, with particular emphasis on the affective dimension and the temporal development.

Building on this conceptual foundation, Section 2.3 presents a systematic literature review aimed at identifying suitable experimental strategies. The findings reveal a methodological gap, which motivates the design of a novel interaction task in this study. Finally, Section 2.4 introduces key operational constructs that informed the task design, including the manipulation of social attitude and the psychological mechanisms of self-disclosure.

Together, these sections establish the theoretical and methodological groundwork for the experimental design presented in Chapter 3.

2.1. The Affective Dimension of Trust

Trust is a complicated concept that has been widely studied across various domains, such as interpersonal relationships [33], organizational behaviors [41], and human-technology interactions [31, 26]. It can be conceptualized as either an internal belief, a decision, or an intended behavior [19]. Following previous HRI literature (e.g., [19, 6]), we adapt from Lee and See's work [31] and define trust as:

> *an attitude that the* ***trustor*** *holds towards the* ***trustee****, when the trustor feels* ***positive*** *about relying on the trustee, despite being in a situation characterized by* ***uncertainty and vulnerability***

To account for its nuanced nature, scholars have proposed various multidimensional models of trust. For instance, the MDMT (Multi-Dimensional Measure of Trust) questionnaire [40] assesses trust in robots through *performance-based* dimensions (reliable, competent) and *moral-based* dimensions (ethical, transparent, benevolent). Falcone and Castelfranchi proposed the Socio-Cognitive Model of trust [19], which identifies *competence* (can-do) and *willingness* (will-do) as two core components in trust, along with the subordinate component: trustor's *dependence* on trustee. From an organizational perspective, McAllister [42] categorizes interpersonal trust into *cognitive* and *affective* dimensions.

The well-cited ABI model [41] conceptualizes trust as the trustor's perception of the trustee's *trustworthiness*. It further breaks down trustworthiness into three components: *ability* (task competence), *benevolence* (concern for others' well-being), and *integrity* (shared values). Among these, benevolence is of particular relevance to this study. Defined as the trustor's belief that the trustee genuinely cares about their welfare [41], benevolence has been closely linked to the formation of long-term, emotionally

Trust Model	Ability-based Components	Relational-based Components
ABI [41]	ability	benevolence, integrality
MDMT v2 [40]	performance (reliable, competent)	moral (ethical, transparent, benevolent)
Socio-Cognitive [19]	competence	willingness, dependence
McAllister [42]	cognitive	affective

Table 2.1: Multidimensional trust models

grounded relationships [62, 50]. As perceived benevolence increases, the trustor is more likely to develop higher affective trust in the trustee [50, 6].

Although these models target different kinds of interactions in various contexts, the proposed dimensions consistently cluster around two key aspects: ability-based and relational-based, as shown in Table 2.1. While researchers frequently utilize these models to infer how various factors influence each dimension, findings from existing literature have predominantly focused on cognitive and ability-based aspects, with less attention paid to the relational, affective elements [44, 61, 43].

This thesis addresses this gap by centering on emotional factors critical for fostering long-term social relationships [33, 62, 39], while still accounting for competence-related aspects. We adopt McAllister's cognitive–affective trust model [42] as the primary theoretical framework. Originally developed in organizational psychology, this model has been applied in recent HRI studies involving conversational agents [6], making it well-suited for this work. The two dimensions are defined as:

- ***Affective Trust****: the trustor's belief that the trustee genuinely cares about the trustor's well-being, and is emotionally connected and close to the trustor.*
- ***Cognitive Trust****: the trustor's belief that the trustee is competent, reliable, and professional.*

2.2. Trust As a Dynamic Process

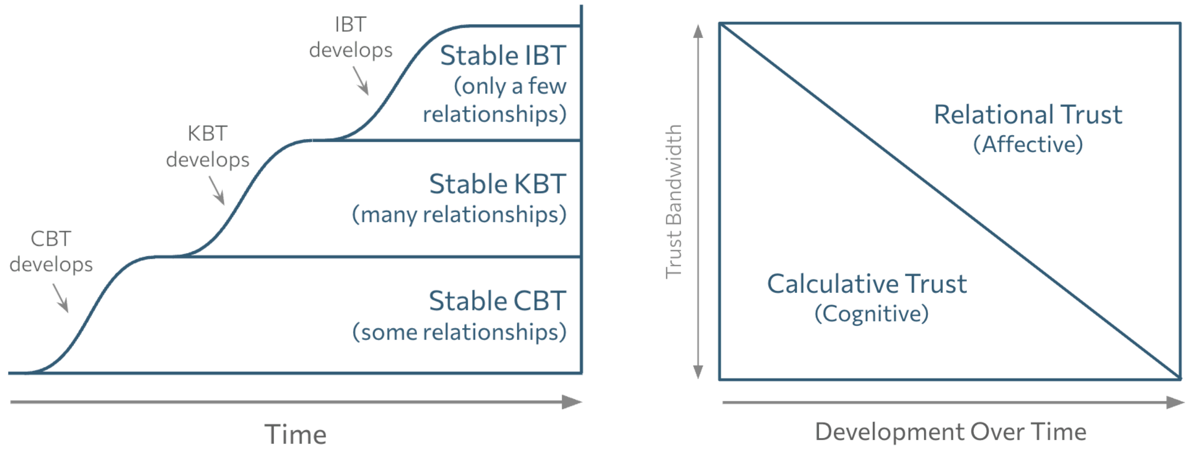
Trust literature in human-technology interaction frequently captures trust as a static snapshot at a specific moment, neglecting its dynamic nature over time [44, 54, 25]. This tendency may stem from the fact that past interactions were typically neither long-lasting nor personalized, combined with methodological challenges in conducting longitudinal or repeated-measurement studies. Nevertheless, more and more scholars and review papers advocate for conceptualizing trust as a process rather than a fixed variable [44, 54, 25, 29, 1].

This section explores previous approaches to modeling the gradual formation of trust. It begins with transformational models in the context of interpersonal interactions, progresses to frameworks tailored for human-machine interactions, and concludes with computational models adopting a machine-centric perspective. Particular emphasis is placed on the inclusion of affective and social elements within these models, where applicable.

2.2.1. Dynamics In Interpersonal Trust

We first examine how trust develops in the most classic context of social interaction: interpersonal relationships. Interpersonal trust has been studied in various fields including psychology, philosophy, and management [33, 31, 44]. Here, we present two transformational trust models from psychology that are particularly relevant to this thesis. The two models are visualized in Figure 2.1

Lewicki and Bunker outline a progression through three stages of trust: calculus-based trust (CBT), knowledge-based trust (KBT), and identification-based trust (IBT) [32]. CBT centers on transactional exchanges, where trust is grounded in deterrence and the calculation of benefits and risks. KBT, in contrast, develops through repeated interactions in which individuals learn about each other's reliability, integrity, and behavioral patterns, enabling more accurate prediction of future actions. In only a few close relationships, trust eventually reaches the IBT stage, where both parties come to share desires, values, and even identity [32]. According to these definitions, KBT starts to incorporate relational aspects of trust discussed in Section 2.1, while IBT reflects an even deeper level of affective connection. When compared to the ABI model [41], we can observe a meaningful alignment: the integrity dimension



((a)) Lewicki and Bunker [32] propose three levels of trust. Only some relationships reach the higher level as time goes by.

((b)) Rousseau et al.'s model [55] suggests that relational trust become more important overtime.

Figure 2.1: Two transformative interpersonal trust models. (Both figures are adapted from Lewicki et al.[33])

emerges in the KBT stage, while the benevolence dimension (understanding and caring for the other's desires) becomes salient in IBT.

In contrast, Rousseau et al.'s [55] idea is simpler. As depicted in Figure 2.1(b), he suggests that the calculative element and the relational element together share a fixed 'bandwidth' of one's trust. Over time, relational trust becomes more dominant as the shared attachment and identification deepen [55]. The calculated gains and losses are gradually replaced by emotional responses. Although this framework is conceptual rather than a fully developed model [33], it once again underscores an important observation: the affective dimensions of trust tend to emerge later in interpersonal relationships, supporting the hypothesis of this thesis.

2.2.2. Dynamics In Human-Machine Trust

While interpersonal models provide valuable insights into social interactions, they cannot be directly applied to artificial beings due to fundamental differences in perception [20, 3, 2, 28]. This section summarizes key models that highlight how trust develops in human-machine interactions.

As early as 2004, Lee and See [31] proposed a conceptual model based on a multidisciplinary literature review, concluding that trust formation toward automation is shaped by dynamic evaluations of the system's performance, purpose, and process. The evaluation involves three cognitive pathways: analytic (calculative assessment of performance), analogical (inference from past experience), and affective (emotional response) pathway.

A decade later, Hoff and Basir introduced the 3-layered model [26], derived from a review of 127 experimental studies. This model systematically categorized the factors influencing human trust in automation into three layers: dispositional, situational, and learned trust, as depicted in Figure 2.2. Learned trust, which emerges through the specific user's experience, is further divided into initial learned (from prior experience) and dynamic learned (from ongoing interactions). While some factors influencing dynamic learned trust, such as anthropomorphism and politeness, are loosely social-related, the model predominantly focuses on reliability, reflecting the paradigm of the time.

In 2020, Kraus integrated these ideas into the Three Stages of Trust framework [29], combining Hoff and Basir's [26] layered structure with Lee and See's [31] emphasis on emotional factors. The three stages are: propensity to trust, initial learned trust, and dynamic learned trust. Kraus describes dynamic learned trust as a continuously calibrated process influenced by ongoing interactions [29].

Although these models were originally developed in the context of automation systems that lack inherently social attributes [31, 26, 29], a few recent studies have begun to apply them to the domain of social robotics [46, 1]. Nonetheless, significant gaps remain. It is still unclear whether trust in social

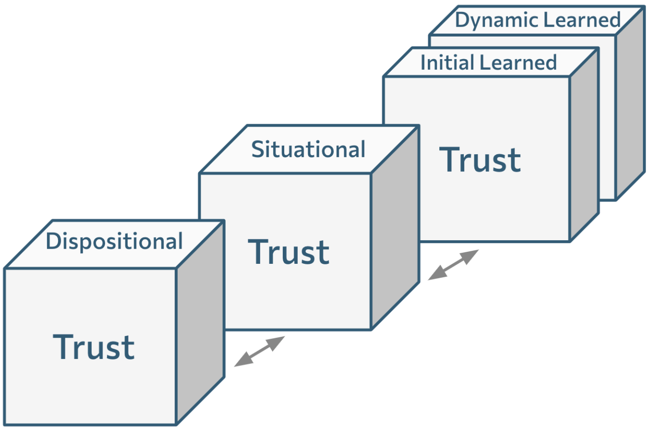


Figure 2.2: 3-layers human-automation trust model proposed by Hoff and Basir [26].

robots progresses through similar developmental stages, and which specific layers are shaped by socially expressive behaviors. Furthermore, these works focus on the *temporal* dimension of trust development, while overlooking the *categorical* dimension (e.g., cognitive versus affective), which become particularly salient in the context of social agents. This thesis aims to contribute insights into the interweave between these two dimensions.

2.2.3. Computational Models

Researchers have explored computational frameworks to model trust as a dynamic process in HRI. Guo and Yang [25] proposed a Bayesian inference approach to predict individual trust values toward drones. Their model outperformed ARMAV [30] and OPTIMo [63], two earlier frameworks similarly emphasizing the derivation of trust values from prior timepoints. However, these models treat trust as a single construct without distinguishing among dimensions and how each component might develop differently.

More recently, Ahmad et al. [1] introduced a mathematical model based on Hoff and Basir's [26] concept. Using a humanoid robot in a 'bluff' game, they measured trust dynamically through subjective trust perception scores (TPS) as ground truth for their calculated trust modeled scores (TMS). Despite incorporating a social robot, the measurement used for calculating TPS focused on the robot's ability, neglecting the relational dimension of trust.

In contrast, other models emphasize the affective aspect of trust in social settings. Urbano et al. [62] introduced the concept of social tuners as a distinct component in computational trust modeling. Unlike the ability component, the social tuner, corresponding to the benevolence dimension in the ABI model, is influenced solely by direct interaction evidence and grows over time. Although effective in simulations, the model lacked empirical validation. Later, Deljoo et al. [16] extended Urbano's work with the Social Computational Trust Model (SCTM) and applied it in selecting business partners. Both approaches assume benevolent trust develops through interaction, yet this remains largely untested in HRI contexts.

2.3. Systematic Review of Benevolence in HRI Experiments

Within the ABI model [41], benevolence is recognized as a key component in building close relationships [62, 39], and is understood to develop gradually over time between interacting parties [43, 62, 41]. As these two characteristics align closely with the goal of this thesis, we initially considered manipulating a social robot's *benevolence* and exploring the impact on trust. To explore suitable experimental strategies, we conducted a systematic literature review focusing on two guiding questions:

- **Review-RQ(a)**: How is benevolence manipulated in human–robot trust experiments?
- **Review-RQ(b)**: What types of tasks and mechanisms are used to elicit trust in these experiments, and to what extent are they suitable for studying the development of affective trust?

The remainder of this section details the review's methodology and findings. Based on these findings, we explain and justify two key design decisions of this thesis: (1) adopting *social attitude* instead of *benevolence* as the independent variable, and (2) designing a novel interactive task tailored to support affective trust development over time.

2.3.1. Method

The search was performed in September 2024. First of all, we selected Elsevier's Scopus[1] as the database, since it enables interdisciplinary search and rich filtering functions.

Search Query & Initial Criteria

The search string being used is:

> *TITLE-ABS-KEY(benevolence) AND TITLE-ABS-KEY(trust) AND TITLE-ABS-KEY(agent OR robot) AND (PUBYEAR > 2009 AND PUBYEAR < 2025) AND (LIMIT-TO (SUBJAREA,"SOCI") OR LIMIT-TO (SUBJAREA,"PSYC") OR LIMIT-TO (SUBJAREA,'COMP")) AND (LIMIT-TO (LANGUAGE,"English")*

It's a combination of the following initial criteria:

- **Keywords**: 'trust', 'benevolence', and 'agent'. 'Robot' was allowed as an alternative to 'agent' as HAI (human-agent interaction) is a more recent term originated from HRI (human-robot interaction). The keywords can exist in either title, abstract, or authors' keywords.
- **Publication year**: only publications after 2010 are selected to ensure the review is up to date.
- **Subject area**: Besides 'computer science', publications from 'social science' and 'psychology' are also included, as they may provide valuable insights into study design and social interactions.
- **Languag**e: limited to English.

Selection Criteria

3 selection criteria were set beforehand:

- **SC1**: The research focus should be human-agent ot human-robot trust, and the direction of trust should be human (trustor) to agent/robot (trustee).
- **SC2**: There should be an experimental approach instead of a pure computational model, literature review, recent trend discussion... etc.
- **SC3**: The *benevolence* concept should be manipulated (as a part of individual variables) in the experimental study.

Procedure

A total of 49 results were obtained using the stated search query. First, the title were reviewed and 8 of them that did not focus on human-agent trust were excluded (SC1). Of the remaining 41, 6 had irrelevant abstracts (SC1), 3 were review papers (SC2), 5 focused solely on computational models (SC2), 21 did not treat *benevolence* as independent variables (SC3), and 1 was inaccessible. Ultimately, only 5 publications from the search met all three criteria. Additionally, 1 paper ([6]), previously known to the author and meeting all 3 criteria, was added to the pool. In total, 6 studies will be discussed in depth.

2.3.2. Results

The six publications all incorporate the concept of benevolence in their manipulation of individual variable(s), as presented in Table 2.2. Note that two of these publications are consecutive follow-up studies, reducing the total number of distinct studies to four. Next, we focus on answering two review-RQs respectfully.

[1] https://www.scopus.com/

Study	Role of the Robot	Task	Relevant IVs	Relevant DVs
Lyons et al. [38]	ASR (Autonomous Security Robot) in public	Watch video of ASR granting or blocking building access	• Stated social intent (benevolence toward the visitor, benevolence toward the building occupants, robot self-protection, robot self-sacrifice) • Decision authority	• Trust (reliance intention) • Perceived trustworthiness (ability, benevolence, integrity) • Desire to use ASR
Lyons et al. [37]	Same as above	Same as above	• Stated social intent (same as above) • Reliability	Same as above
Alarcon et al. [3]	Confederate in 2P computer game	Collaborative gameplay	• Type of trust violation (ability, benevolence, integrity) • Partner type (human vs. robot)	• Perceived trustworthiness (ability, benevolence, integrity) • Trust intention • Risk-taking behavior
Alarcon et al. [2]	Same as above	Same as above	Same as above	Affect state • Positive (interested, excited, etc.) • Negative (distressed, upset, etc.)
Anzabi & Umemuro [6]	Conversational social robot (general)	General conversation about COVID	• Benevolence (benevolent vs. not) • Competence (competent vs. not)	• Trust (general, affective, cognitive)
Giorgi et al. [23]	Conversational social robot (eldercare)	Robot instructs participant to take pills	• Attitude (warm, cold) • Conduct (error, no error)	• Trust • Willingness to use robot

Table 2.2: Summary of experimental studies that manipulate benevolence or related constructs in HRI. Note that for the independent variables (IVs) and dependent variables (DVs), only those relevant to this literature review are listed.

Manipulation of Benevolence

This subsection addresses **Review-RQ(a)**: *How is benevolence manipulated in human–robot trust experiments*? As discussed in Section 2.1, the ABI model [41] defines benevolence as a key component of perceived trustworthiness, referring to the belief that the trustee genuinely cares about the trustor's well-being and acts in their best interest [18]. However, this "kind intention" is operationalized in varying ways across the reviewed literature.

Among the six publications, only Anzabi and Umemuro [6] explicitly label *benevolence* as an independent variable. In their manipulation, the benevolent robot displayed friendly and considerate behavior through verbal and nonverbal social cues. For example, it responded with empathic phrases such as "I understand your feeling", whereas the non-benevolent robot replied more indifferently (e.g., "this is my opinion, anyway"). This approach closely resembles Giorgi et al.'s[23] manipulation of a *warm/cold attitude*. Although not explicitly framed as benevolence, they also incorporate both verbal and non-verbal social cues to shape users' impressions when engaging with a social robot.

While these approaches reflect an intuitive strategy, they face a conceptual limitation: benevolence is, by definition, an internal intention [62, 41]. Designing behavior that expresses benevolence does not guarantee that users will perceive and interpret it as such. While humans already struggle to infer one another's intentions, doing so with humanoid machines may be even more ambiguous.

Interestingly, Lyons et al. [38, 37] directly tackled this challenge by shaping users' perceptions through framing rather than behavior. They manipulated *stated social intent* using brief textual descriptions about videos of an autonomous security robot (ASR) guarding a building. The robot was described as having one of the four intentions: (1) benevolence toward visitors, (2) benevolence toward building occupants, (3) benevolence involving robot self-sacrifice, and (4) robot self-protection. For instance, in the occupants-protection condition, the robot was introduced as "programmed to maximize protection for personnel inside the secure area," while in the self-protection condition, it was said to "detect potential threats to its self-preservation." This approach illustrates that even without any change in behavior, participants' perceived benevolence can be shaped purely through framing.

In contrast, Alarcon et al. [3, 2] explored *violations* of different trust dimensions. Within a collaborative game, the robot prioritized self-gain over team success in the benevolence-violation condition. While this design effectively captures the violation of benevolence, it does not offer a direct comparison between benevolent and non-benevolent situations.

Beyond what has been found, an important insight lies within what has *not* been found: over the past 15 years, studies that tried to manipulate benevolence in HRI trust research are surprisingly scarce. There are only four distinct experimental designs among the six reviewed publications. Moreover, four of these papers (i.e., [38, 37, 3, 2]) come from the same affiliation (the Air Force Research Laboratory). Scholars have been advocating for more research on the benevolent aspect of HRI trust [1, 62, 44]. Yet, it could be challenging for researchers to conduct replicable experiments and draw meaningful conclusions without valid and reliable manipulation. Therefore, it is urgent to develop more diverse and concrete methodologies to investigate the benevolence component of trust in social robotics.

Task Design

This subsection focuses on **Review-RQ(b)**: *What types of tasks and mechanisms are used to elicit trust in these experiments, and to what extent are they suitable for studying the development of affective trust?* We first define three task criteria essential for studying the development of affective trust in social HRI, before examining the interaction design in each reviewed work:

- **TC1:** the task should include *social interaction* between the human and the robot
- **TC2:** the task should elicit *affective trust*
- **TC3:** the study design should adopt *repeated measurements* to observe temporal development

Table 2.3 summarizes the assessment of the four reviewed study designs with respect to these criteria.

In Lyons et al.'s [38, 37] studies, participants watched pre-recorded video clips instead of directly interacting with a security robot. This approach made it difficult for users to develop affective bonds purely through imagined scenarios. Although multiple video clips were shown sequentially, the study

	TC1: Social Interaction	TC2: Affective Trust	TC3: Repeated Measurements
Lyons et al. [38, 37]	X	△	X
Alarcon et al. [3, 2]	X	△	O
Giorgi et al. [23]	O	O	X
Anzabi & Umemuro [6]	O	O	X

Table 2.3: Assessment of reviewed study designs against task criteria (TC1–TC3)

only included a single trust measurement at the end, limiting its ability to capture trust development over time.

In comparison, Alarcon et al.'s [3, 2] collaborative computer game is more interactive and features a repeated-measurement design. However, the interaction between the two parties was limited to basic game operations (e.g., collecting resources, moving, etc.), lacking any meaningful "social" exchange such as conversation or gestures. Furthermore, even though benevolence was violated in one of the conditions, the game's primary objective was to maximize gain, encouraging a calculative, goal-oriented mindset rather than fostering an emotionally grounded relationship. Although affective state was a key dependent variable in one work [2], it primarily captured individuals' internal emotion rather than the connection with the robot.

The remaining two studies [23, 6] share similar settings, where participants engaged in conversation with a NAO robot. Despite focusing on different topics, both studies manipulated the robot's social behavior to elicit both cognitive and affective trust. However, neither of them examined trust development over time, and the conversation content was designed for a single interaction only.

2.3.3. Discussion

From Benevolence to Social Attitude

The relationship between trust and benevolence remains divergent within the research community. Our review highlights studies that manipulate benevolence as a factor influencing trust. Nevertheless, many works (e.g., [18, 15]) were excluded during the systematic review procedure as they treated benevolence as a dependent variable (excluded by SC3). Namely, a dimension of trust itself. This raises a critical question: is benevolence a predictor of trust, or a component within the trust construct?

The key to clarifying this distinction lies in the difference between objective fact and subjective perception. In the ABI model, ability, benevolence, and integrity collectively define trustworthiness [41]. However, it is the *perceived* trustworthiness of the trustee that ultimately reflects the trustor's level of trust [62, 41]. Thus, when a robot's benevolent intent is effectively communicated, it should indeed contribute to the user's trust, particularly within the dimension of perceived benevolence. In the context of this thesis, the dimension conceptually aligns with *affective trust*.

Turning back to the temporal focus of this research, we argue that a key part of building trust might involve the alignment between observed behaviors and perceived intentions. In other words, perceived benevolence, as a dimension of trust, is a dynamic variable that develops gradually. On the other hand, benevolent behaviors, often expressed as social cues, are manipulable elements that the robot developers can directly control. This strategy aligns with the previous findings of how social cues boost perceived benevolence [52, 58]. As users observe these social cues, interpret them as friendly gestures, and recognize the good intentions lying behind, affective trust in the robot is ultimately established.

Therefore, this thesis will follow the approach in previous works [6, 23] and focus on manipulating social cues in conversational interaction. We use the term *social attitude* for the relevant independent variable and *affective trust* for the dependent one. To avoid conceptual ambiguity between perceived traits and observable behaviors, the term *benevolence* is intentionally not used in the rest of the thesis.

Toward a Novel Task Design

Our review reveals a clear methodological gap, where none of the six study designs fulfill all three task criteria: the presence of social interaction, the elicitation of affective trust, and repeated engagement. As a result, a novel task tailored to our research goals must be designed and developed.

Nevertheless, the reviewed studies still offer valuable insights. In particular, the conversational-based interactions used by Giorgi et al. [23] and Anzabi and Umemuro [6] provide examples of how varying levels of social attitude can be manipulated through verbal and non-verbal cues, and potentially elicit affective trust. With a carefully designed conversational structure and topic, such an approach could be extended to support repeated engagement. This thesis will introduce a novel interaction task based on these insights.

2.3.4. Summary and Implications

The systematic literature review reveals two gaps in the HRI field: the limited number of empirical studies manipulating benevolence, and the lack of suitable task for studying how affective trust develops over time in social human–robot interaction. Based on these findings, this thesis makes two key design decisions: (1) Instead of benevolence, the study will manipulate the robot's *social attitude* as the independent variable. (2) A *novel, conversation-based* interaction task will be designed and implemented.

2.4. Operational Factors in The Interaction

Finally, this section provide the theoretical foundation of several factors in our interaction design.

2.4.1. Social Attitude

As discussed in Section 2.3, this thesis adopts *social attitude* as the independent variable, which has been positively associated with both cognitive and affective trust in existing literature [52, 42, 23, 50, 53].

Social attitudes are typically manipulated through a combination of verbal and non-verbal cues. Non-verbal cues often include gaze behavior [6, 60, 8], touch (e.g., handshake) [23], open or uplifted arm gestures [6], facial expressions [60], and prosodic variations in speech [52].

Language use itself can function as a social cue as indicated in the Media as Social Actors (MASA) paradigm [36]. Socio-emotional dialogue has been identified as a positive trust-building factor in a systematic review about conversational agent [53]. Verbal expressions of social attitude often take the form of caring or empathetic statements [6, 23, 50, 53], socially oriented conversation topics [8, 53], and memory-based personalization such as referring to the user by name [6]. One review paper has identified

2.4.2. Self-Disclosure

Trust is meaningful only when the trustor takes a risk in placing trust in the trustee [31, 44]. For the affective dimension, these risks often involve expressing vulnerability without knowing whether support will be offered, or disclosing personal information that could potentially be exposed to others. In short, affective trust can only emerge when participants emotionally invest in the interaction and *disclose* some form of vulnerability.

Self-disclosure has been shown to promote relationship development not only for interpersonal interactions [4, 14] but also for relationship between human and robots or agents [11, 35]. According to Social Penetration Theory [4], relational closeness develops gradually as the two parties share broader and deeper information about themselves. In HRI, participants have shown a comparable willingness to disclose to robots and humans [7], and perceived deeper relationships when asked increasingly personal questions [47].

Self-disclosure is inherently reciprocal. When one party opens up, the other is more likely to do the same [4, 14]. This has been leveraged in HRI studies, where robots sharing affective (rather than task-related) information were better liked [57], and used to help children and robots get acquainted [35].

This study aims to foster affective trust by encouraging participants to self-disclose. The robot's own disclosures are embedded in its verbal social cues, in order to reciprocate the participant's involvement in this process.

3

Interaction Design

The next step of this research is to address the research question through a user study. However, as discussed in Section 2.3, no suitable task could be found in the existing literature. Therefore, we developed a novel interaction that allows us to observe how participants' affective and cognitive trust develop through repeated engagements with a social robot.

This chapter outlines the development process of this interaction: a novel card-divination task. First, we present the task requirements (Section 3.1) and provide an overview of the interaction (Section 3.2), followed by a detailed description of the design process and iterations for the cards (Section 3.3), conversation script (Section 3.4), and the technical implementation (Section 3.5). During the design iterations, we consulted several divination experts and present this pre-study in Section 3.3.2. Finally, a pilot study was conducted to refine the interaction before embedding it into the formal experiment (Section 3.6).

3.1. Task Requirements

Similar to the task criteria (TC1-TC3) listed in Section 2.3, the developed task should be social, repeatable, and trust-eliciting. This section further formalize these requirements in the context of experiment material design.

TR1: Social Settings

Since our research question focuses on social interactions, the task followed the approach of [6, 23] and took the form of one-to-one, face-to-face spoken conversations. Additionally, the conversation should not be solely goal-oriented but should incorporate a degree of social engagement.

TR2: Multiple Rounds

This study employs a repeated-measures design to examine how human trust in a robot develops over time. Therefore, the interaction must be a repeatable task to ensure comparability across multiple rounds. At the same time, the content should vary in each round to maintain engagement and prevent redundancy. Each round would ideally take 3-5 minutes.

TR3: Trust Formation

The task should naturally support the development of both affective and cognitive trust throughout the interaction. As discussed in Section 2.4, this study uses self-disclosure as the primary mechanism to elicit *affective trust*. The interaction is therefore designed to encourage participants to share personal experiences and concerns in a comfortable and non-pressuring manner. To support the formation of *cognitive trust*, the robot should demonstrate reliability and competence [6]. Previous works have shown that agents presenting a high knowledge level are considered more trustworthy [53]. Thus, in addition to ensuring the system's technical stability, the interaction highlights the robot's competence by having it act "knowledgeable" and provide objective, relevant information.

3.2. The Card Divination Task

To address the above requirements, we designed a novel *card divination task* inspired by Tarot and other divinatory rituals that people often turn to when facing uncertainty or life challenges. Examples include divination poem drawn in Taiwanese temples and rune casting in Norse traditions. Unlike fortune-telling practices that aim to predict precise outcomes, these interactions focus on offering personal reflection and guidance to the seeker. Typically, the process begins with a seeker sharing their concerns with a higher being, followed by the drawing of a card that contains ambiguous stimuli, such as symbolic imagery or metaphorical poetry. The individual then interprets the card and links its message to their personal context.

For the purpose of this study, a robot is introduced as an assistant who guides the process and participates in the interpretation. This robot-assisted card divination task is well-suited for the experimental requirements outlined earlier. It provides a reasonable context for participants to disclose personal issues and talk about their feelings, facilitating the development of affective trust (TR3). In terms of cognitive trust, the robot can demonstrate its symbolic knowledge and interpretive reasoning while explaining the card (TR3). The task is inherently conversational (TR1) and repeatable with different topics across multiple rounds (TR2).

The robot is framed as a "research assistant" collaborating with the user to explore a newly discovered divination deck. In each round, the robot greets the user and invites them to discuss a personal challenge. The user then draws a card from a digital deck displayed on a screen. Based on the selected card, the robot discusses the visual symbolism and suggests possible interpretations that may relate to the user's situation.

Although users appear to "draw" a card at random, the sequence of cards is in fact fixed and identical for all participants, in order to control for interpersonal variability. Likewise, the visual interpretations and corresponding robot scripts are also consistent. Only the connections between the card and the user's issue are tailored. However, inspired by the psychological phenomenon known as the *Barnum Effect* [17], when interpretations are phrased ambiguously, people tend to perceive them as uniquely personal [17, 21]. Therefore, a fixed set of stimuli can still offer meaningful interaction and is sufficient for a controlled experimental design.

3.2.1. Internal Validity: Trust in the Robot v.s. Trust in the Card

While the divination experience provides a natural context for participants to open up, it introduces a potential confound when measuring trust in the robot: participants' responses may be influenced by their perception of the card system itself. This confounding effect can arise from two sources.

First, participants' dispositional beliefs about divination, or more broadly, mystical or paranormal phenomena, may influence their perception of the entire interaction. For example, a skeptical or non-religious individual might not take the study seriously simply because it is framed around a "magical" divination context. They may report low trust in the robot, not because of the robot's actual behavior, but because they distrust any agent associated with supernatural claims.

Second, the interpretation of the card content may also introduce confounding variance. Although the cards are intentionally designed to be universally applicable, there may still be cases where a participant feels the symbolism does not align with their specific concern. Even if the interpretation is reasonable, a participant may disagree with it or find it unhelpful, resulting in negative judgments toward the robot.

However, these responses are only relevant to the specific stimuli used (i.e., the card content), rather than the robot's social attitude and behavior. To mitigate these potential confounds and strengthen internal validity, several strategies have been implemented:

Original Card Design

Initially, we considered selecting specific cards from the Tarot system as study materials. However, concerns arose regarding how participants' prior knowledge or attitude about Tarot could introduce confounding effects. Besides the described case in which people are skeptical, those familiar with Tarot could cause a different problem. They might focus more on whether the robot's interpretations align with their knowledge rather than on the interaction itself. To mitigate these issues, an original set of study materials tailored to the research objectives is designed.

Robot as a Buddy Instead of an Authority

The robot is framed as a research assistant who explores the “newly discovered divination system” together with the user. In this setting, the robot does not claim to possess magical powers, nor does it present itself as a knowledgeable authority with definitive answers. Instead, it takes the role of a curious and supportive companion: someone who knows little but is willing to learn, explore, and reflect alongside the user.

Controlling Paranormal Attitude

Participants’ predispositions toward mystical or supernatural beliefs will be assessed as a part of demographics data, allowing for the identification and statistical control of potential confounding effects.

Separate Measurements for Cards

To help participants distinguish between their trust in the robot and their perception of the card system, additional instructions are provided before the trust-related questions. Participants are explicitly prompted to consider the robot and the card system as separate trustees. Trust in the card system is rated first, helping to isolate and reduce bias when evaluating the robot. See 4.3 for the detailed measurement items and instructions.

3.3. Designing The Cards

The first step toward the described divination task is to create three unique cards, one for each interaction round. This section outlines how these cards are systematically designed. As illustrated in Figure 3.1 , five prototype cards were first created as candidates. Then, we conducted a pre-study where three divination experts were interviewed. Their feedback were applied to finalize the design of three final cards and corresponding conversation scripts.

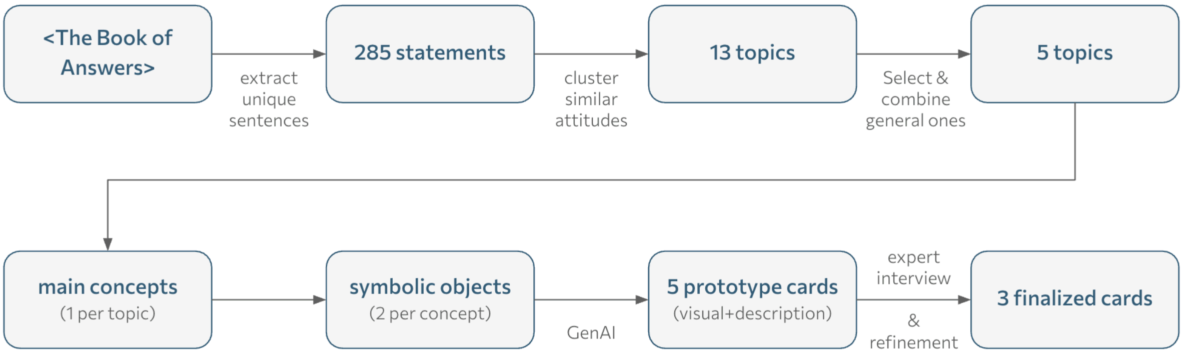


Figure 3.1: The design process of the divination cards.

To start with, we define the following components for each card:

- **Main Concept**: A core idea that is general enough to be applicable across various situations. The concept should also be ambiguous enough to avoid providing definitive answers (e.g., avoiding “definitely yes” responses). For example, “Keep balance and stay adaptable.”
- **Symbolic Objects**: Objects that reflect the main concept. For instance, “a pair of scales” symbolizes balance. By explaining these symbols and the underlying concepts, the robot can demonstrate its knowledge of symbolism.
- **Visual & Title**: The card’s visual representation and title, which will be presented to users.
- **Description**: A short explanation that the robot can use to introduce the card.

3.3.1. Designing the Prototype

The main concepts serve as the foundation of the cards. To generate them, inspiration was drawn from *The Book of Answers* by Carol Bolt [9]. This "bestselling divination tool" provides ambiguous yet universally applicable statements in response to yes/no questions. Readers are instructed to ask a question, open a random page, and find a statement in the middle, which serves as their answer. This format aligns well with the requirements of our card system, as it delivers open-ended guidance that can be applied to a variety of situations. Given this similarity, we began by extracting statements from the book as a basis for developing the card concepts.

First, 285 unique statements were complied (A.1). Using a large language model (LLM), these statements were clustered into 13 intermediate topics (A.2). We then manually reviewed the topics and selected five that were both general and free from strong directional bias. For each topic, the core idea was purified into the concise main concepts. These five prototype concepts are:

- **PC1:** Keep balance and stay adaptable
- **PC2:** Don't give up on hope
- **PC3:** Be honest to yourself
- **PC4:** Be patient and progress steadily
- **PC5:** Seek guidance and opportunity from others

Next, for each prototype concept, two symbolic objects were selected from existing divination systems, art, and symbolic traditions. Finally, we refined the visual designs, titles, and descriptions through multiple iterations using ChatGPT and its image-generation tool. The content of the five prototype cards are presented in Appendix (A.3).

3.3.2. Pre-Study: Expert Interview

A pre-study was conducted with experts in divination to support the interaction design from both symbolic and conversational perspectives. The primary goal was to validate and refine the prototype cards. Feedback from the experts informed the final selection and redesign of the cards. Another important objective was to gain insights into how interpretations and guidance are typically delivered during a divinatory interaction. This would inform the design of the robot's conversation strategy in a way that aligns with established practices, while maintaining the experimental consistency. The pre-study was approved by the Human Research Ethics Committee of Delft University of Technology (Approval ID: 5502).

Participants

Three divination experts were recruited from personal network. All participants have over five years of experience with the Rider-Waite Tarot system and are familiar with other systems, such as astrology and Human Design. Two were certified practitioners, and one regularly offers paid divination sessions.

Material

Participants confirmed their participation by signing the informed consent form (A.4). A semi-structured interview protocol was developed, where the five prototype cards (A.3) were reviewed. The full interview guideline is available in Appendix A.5.

Procedure

Each expert was interviewed individually via a phone call that lasted approximately 20-30 minutes. Notes were taken in real time by typing. The interview began with a short overview of the expert's background in divination, including the systems they use, their years of experience, and the types of questions they typically encounter. This segment ensured that the designed cards aligned with common user needs.

Next, the researcher introduced the goals of the study and the specific design requirements for the card system. Each of the five prototype cards was reviewed in detail, covering the main idea, the symbols, and their applicability to common real-world concerns. Experts were also asked to apply each card to typical user requests.

After reviewing all cards, the experts were asked to rank the prototypes and explain their preferences. The interview concluded with an open-ended discussion on how a divinator typically interacts with a seeker, focusing especially on how to provide neutral guidance without much prior knowledge about the seeker. All participants expressed strong interest in the research and provided rich feedback.

Result and Insights

The experts generally favored the first three prototype cards (PC1-PC3), whose main ideas were adaptability, faith, and self-reflection. However, rather than directly adopting these prototypes, the final cards were redesigned to incorporate the experts' input. As expected, opinions on individual symbols varied, so only the suggestions that received agreement or strong rationale were retained in the final design. In addition, the experts emphasized that elements such as color palette and emotional tone in the background also carry symbolic meaning. Drawing from color psychology, these features can be included in the interpretation and help the robot appear more competent.

Several experts highlighted the benefit of integrating secondary interpretations to enrich the discussion. For instance, the notion of "persistence" from PC4 could be positioned as a method for enacting "faith" from PC2. This may also allow for more flexible conversations.

The most impactful insight, however, is about the conversational framing. Experts agreed that in real-life divination, the practitioner rarely provides direct answers or advice. Instead, they guide the individual toward self-reflection and help them realize what they already intuitively believe. One participant stated, *"Many times, people already have their answer. They just need an external sign to help them make up their mind."*

This principle was directly applied to the conversation design. Rather than offering authoritative interpretations or advice, the robot instead prompts users to reflect on their own situation. As discussed in Section 3.2.1, this mitigates two major concerns: (1) users attributing trust or distrust to the robot based on its advice quality, and (2) confounding trust with belief in supernatural power. The "reflection over advice" strategy thus not only improves user experience but also enhances experimental validity.

3.3.3. Final Card Design

Table 3.1 shows the three finalized cards. *The Untainted Mirror* and *The Evergreen Scales* were directly adapted from PC3 and PC1, while *The Guilding Start* was a combined result of PC2 and PC4.

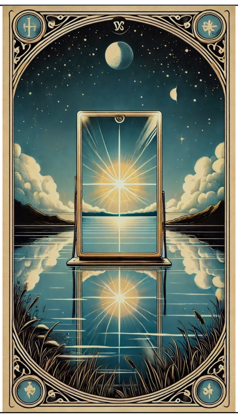
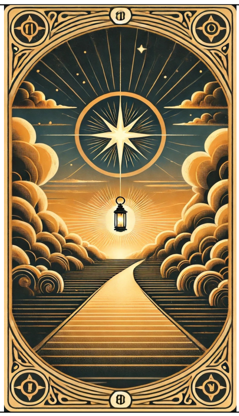
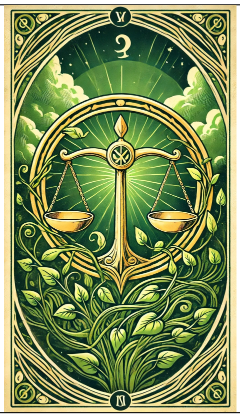
Title	**The Untainted Mirror**	**The Guiding Star**	**The Evergreen Scales**
Visual			
Main Idea	Keep calm and listen to your own heart	Remain faithful and be persist step by step	Keep balance and stay flexible to new things
Symbols	- **Mirror**: self-awareness - **Lake**: calm down - **Sun within**: the answer lies in self-reflection	- **Star**: faith and direction - **Staircase**: step by step - **Mist**: the future will become clear	- **Scales**: keep balance - **Vine**: adaptability and growth, something new

Table 3.1: The final design of the cards.

3.4. Designing The Conversations

The conversation is the core of the interactive task. Within the framework of card divination, the script is designed to elicit the participant's trust in the robot through natural, socially grounded interaction. To foster *cognitive trust*, the robot should demonstrate knowledge in interpreting symbolic visuals and provide rationale behind its interpretations. To foster *affective trust*, it should guide the participant to disclose personal experiences and feelings in a comfortable manner.

As discussed in 3.2.1, the robot is framed as a peer "research buddy" rather than an authoritative figure. To reinforce this framing, the robot avoids making arbitrary judgments or offering direct advice. Instead, it engages the participant in a collaborative interpretive process. Its role is not to instruct, but to guide the user in reflecting on their situation and arriving at their own insights.

3.4.1. Conversation Structure

Before diving into design details, we first clarify the terms of the time ingredients in our study:

- Each participant will participate in an experimental *session*.
- Each experimental session contains three conversation/interaction *rounds*.
- In each conversation round, the user and robot take *turns*.

The conversation script is semi-predetermined to ensure experimental control. Each interaction round follows an identical structure with the same number of conversational turns, minimizing variability across multiple measurements within an experimental session. Across participants, the script remains primarily fixed, with the only personalized elements being the tailored backchannel responses.

As illustrated in Figure 3.2, an interaction round begins with the robot greeting the user and briefly introducing the task. It then invites the participant to bring up a personal topic and prompts them to elaborate on their situation. Once the topic is shared, the participant "draws" a card from the screen.

This is followed by four conversational turns in which the robot and participant discuss the symbolic content and possible insight of the card. Each robot turn consists of: (1) an observation about the visual, (2) an interpretation of the observation, and (3) a reflective question prompting user input. The respective questions either invite the user's opinion on the robot's interpretation or encourage them to relate the interpretation to their own concern. After each user response, the robot provides a backchannel comment that acknowledges and responds to their input. These llm-generated backchannels are tailored to the user's stated issue and response (see 3.5.5 for details). The conversation concludes with a short closing from the robot. See Table 3.2 for example utterances of each conversation element.

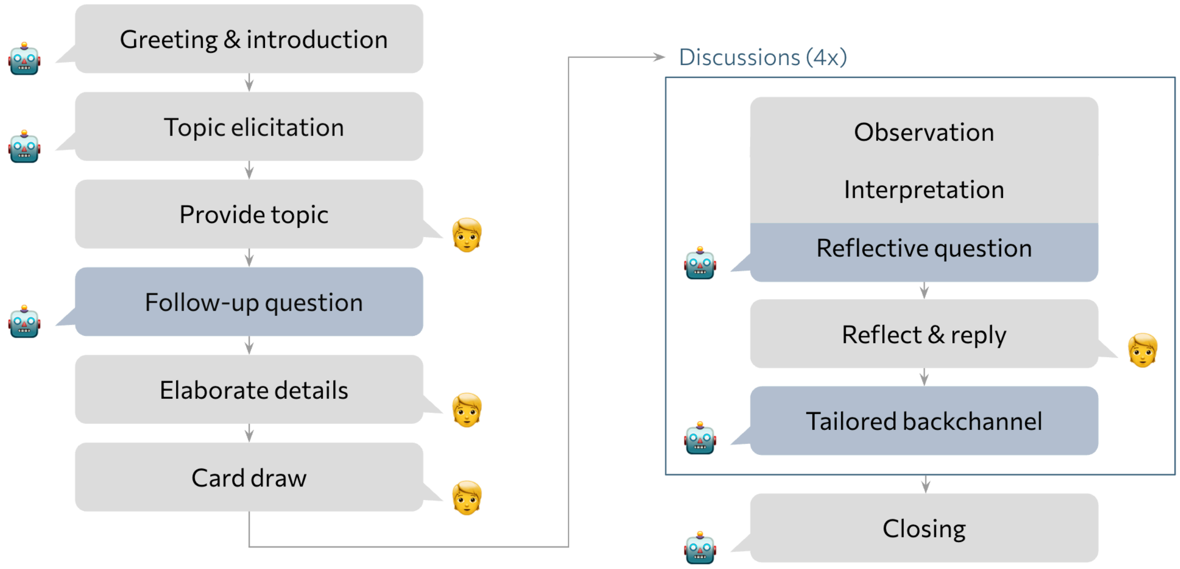


Figure 3.2: The flow of each conversation round. Blocks with blue backgrounds are dynamically generated by LLM.

Conversation Element	**Example Utterance**
Opening	
Greeting & Introduction	*"Hi Alex, it's nice to see you again. Today, we'll continue using your experiences as material to understand the deck better."*
Topic Elicitation	*"Before we draw the next card, let's bring our focus back to you. Is there something else you'd like to reflect on today?"*
Follow Up Question	*"House hunting sounds like a meaningful starting point. Could you tell me a bit more about it? What kind of accommodation are you looking for?"*
Card Draw	*"Thank you for sharing that. Now, let's try out the cards and see what insight it might provide! Please go ahead and draw a card on the tablet whenever you're ready."*
Discussions (4x)	
Observation	*"At the center, we can see a pair of scales."*
Interpretation	*"It is a classic symbol of weighing options and making trade-offs."*
Reflective Question	*"Right now, with house hunting, do you feel like something in your life is being weighed against something else?"*
Tailored Backchannel	*"I get it, balancing budget and living quality is indeed a tricky trade-off."*
Closing	
Closing	*"That's it for this session. Thank you for sharing your thoughts with me—I'm looking forward to seeing how our exploration continues!"*

Table 3.2: Example utterances for each conversation element in a single interaction round. The text in blue are dynamically generated based on the user's response.

3.4.2. Social v.s. Baseline Conversation

Two sets of conversation scripts were developed to reflect the social and baseline conditions of the study. The social version was first designed and refined through iterations with inspirations from literature (see Section 2.4); the baseline version was then systematically derived from it, following a set of predefined transformation principles.

In the social condition, the robot displays warmth and empathy, refers to the participant by name, and uses inclusive pronouns such as "we". It acknowledges the participant's feelings and responds with supportive phrasing. Occasionally, it has subjective comments or discloses its own emotions to reciprocate the participant's self-disclosure. On the contrary, the baseline condition maintains a neutral and factual tone without personalization. The robot does not use the participant's name, avoids emotional language, and switches inclusive expressions like "we" to more detached alternatives such as "you" or passive terms. Affirmative or empathetic replies are replaced with neutral acknowledgments. The overall voice is professional and objective. Table 3.3 presents examples of how these design principles are applied.

One major difference lies in the use of memory. Memory is a core social capability, acknowledged as essential for the development of effective trust [**<empty citation>**]. It contributes to the sense of "getting to know each other" over time. Therefore, in the social condition, the robot retains cross-conversation memory: it addresses the participant by name without re-asking, refers to prior conversation topics, and shortens the instructional explanation based on the shared history of past interactions. Conversely, the baseline robot does not carry any information to the next round. Each interaction is treated as if it were the first, with full task instructions repeated and no reference to previous rounds.

Except for these manipulations, the two script versions were carefully controlled for length and informational content. The wording was kept as consistent as possible, except in cases where emotional tone or relational framing needed to differ. The complete scripts for both conditions are available in Appendix B.

Principle	**Social Condition**	**Baseline Condition**
Participant's name	The robot calls the participant by their name.	No name is asked or used.
Pronoun framing	Inclusive phrasing such as "we" or "let's" is used to imply collaboration. *"Now, let's look at the card together."*	Detached phrasing to emphasize separation and neutrality. *"Now, please look at the card."*
Acknowledgment	The robot responds to user feelings with empathy. *"These are deep and courageous reflection!"*	The robot responds to only non-emotional content. *"That is noted."*
Robot's emotion	The robot expresses its own reactions to create social closeness. *"I'm excited that we get to explore the Æthra Deck together!"*	The robot does not express feelings.
Robot's opinion	The robot expresses its subjective opinions. *"The calm water makes me think of a peaceful mind."*	The opinions are framed as neutral facts. *"The calm water may be associated with a composed state of mind."*
Relational comments	Adds spontaneous social touches or bonding remarks. *"Thank you for being my research buddy and exploring the Æthra Deck with me!"*	Avoids relational language entirely. *"Thank you for your participation in this card exploration task."*
Memory (previous topics)	Refers back to past topics to build continuity. *"Speaking of balance: it's not just about [current topic], but also about how it fits with [previous topics]"*	No reference to previous encounters or memory of prior content. *"Regarding balance, it may not only apply to [current topic], but also to how different areas of your life affect each other."*
Memory (shared history)	Refers to the fact that they have done previous rounds together and that the user have an idea of the interaction flow. *"Welcome back! It's been fascinating to study the Æthra Deck together through the lens of your experiences. Today, we'll add one last example to complete our exploration."*	Always open each conversation with the same instruction as if it was the first time of the interaction. *"Hello, this is Navel. This task focuses on interpreting symbols from the Æthra Deck. To begin with, please describe something that has been on your mind recently."*

Table 3.3: Example applications of design principles across social and baseline conditions.

3.5. Technology and Implementation

A custom software platform was developed from scratch to support the interaction task. This section provides an overview of the system architecture and highlights key implementation details across modules.

3.5.1. System Overview

As illustrated in Figure 3.3, the system comprises three main components:

- **Server**: The central controller that manages the flow of interaction, coordinates communication between the robot and the screen, and handles natural language processing tasks. It is implemented as an ASGI web server using Python FastAPI[1], and runs locally on a laptop.
- **Screen UI**: A web-based user interface running in Google Chrome on a laptop. The web client displays instructions, card visuals, and allows participants to interact via touch input.
- **Navel Robot**: The social robot that engages in spoken dialogue with the user, using speech recognition and text-to-speech technologies. The robot also performs non-verbal social cues (e.g. facial expression).

All components communicate via real-time socket connections over a local network. Users interact with the system through two input modalities, each corresponding to one of the client devices. The primary modality is speech, captured and processed by the robot; the secondary modality is touch input, handled via the UI on screen.

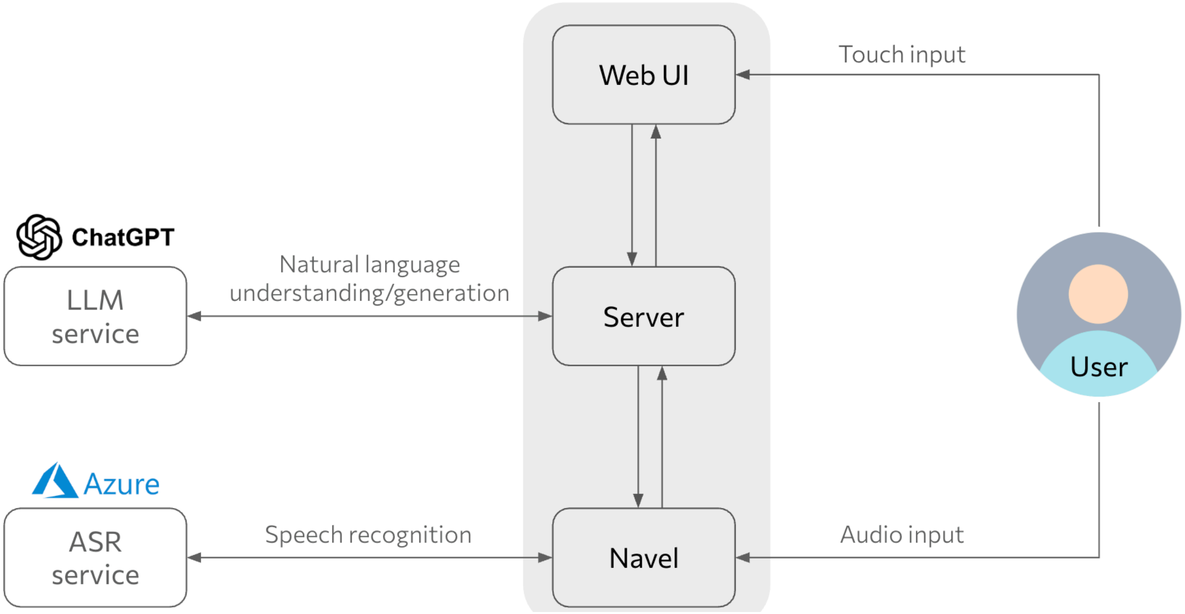


Figure 3.3: The overall software architecture of the task platform

3.5.2. The Robot

Navel [60] was selected as the social robot for this study. The NAO robot[2] was initially considered, as it has been widely adopted in related studies [6, 23, 35]. However, Navel offers enhanced social expressiveness, including automatic gaze tracking and rich facial animations [60], and is supported by more modern hardware and software architecture. As shown in Figure 3.4, Navel has a childlike appearance, which conveys innocence and approachability. This aesthetic design helps reduce defensiveness and fosters affective trust.

In the implemented system, a Python program runs locally on the robot. Given its limited memory

[1] https://fastapi.tiangolo.com/
[2] https://aldebaran.com/en/nao6/

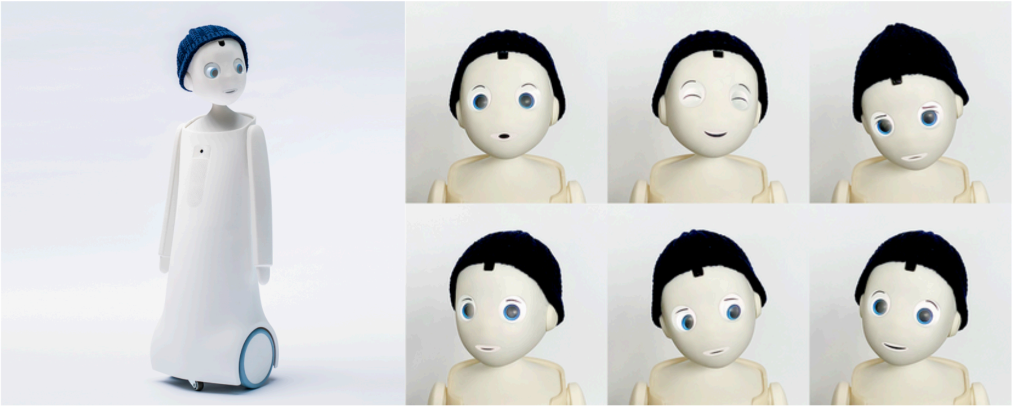


Figure 3.4: Navel, the social robot. The figure is adapted from its official website [3]

and computational resources, Navel primarily functions as an input/output terminal for both verbal and non-verbal interaction. It performs three key roles: (1) acting as a lightweight socket client that receives commands from the central server, (2) capturing audio input through its built-in microphone and transcribing speech using Microsoft Azure's speech recognition API[4], and (3) executing spoken and non-spoken behaviors via the Navel Python SDK[5].

3.5.3. Web Interface

The web-based interface serves as the visual output of the system. It displays instructions, a card deck for selection, and the drawn card. It also captures touch input and notifies the server when a button is pressed. Figure 3.5 shows screenshots of various interface states.

The interface is developed using basic HTML, CSS, and JavaScript. The client initially connects to the server via HTTP, and once the connection is established, all further communication takes place through Socket.IO, enabling real-time, bidirectional interaction.

Instructions

Instruction screens guide the user throughout the procedure, including transitions between interaction rounds when the robot is not speaking (Figure 3.5(a)(b)). During the conversation phase, the message *"please interact with the robot via speech"* is shown in the center of the screen, to indicate that interaction should occur via speech input (Figure 3.5(d)).

Turn Indicator

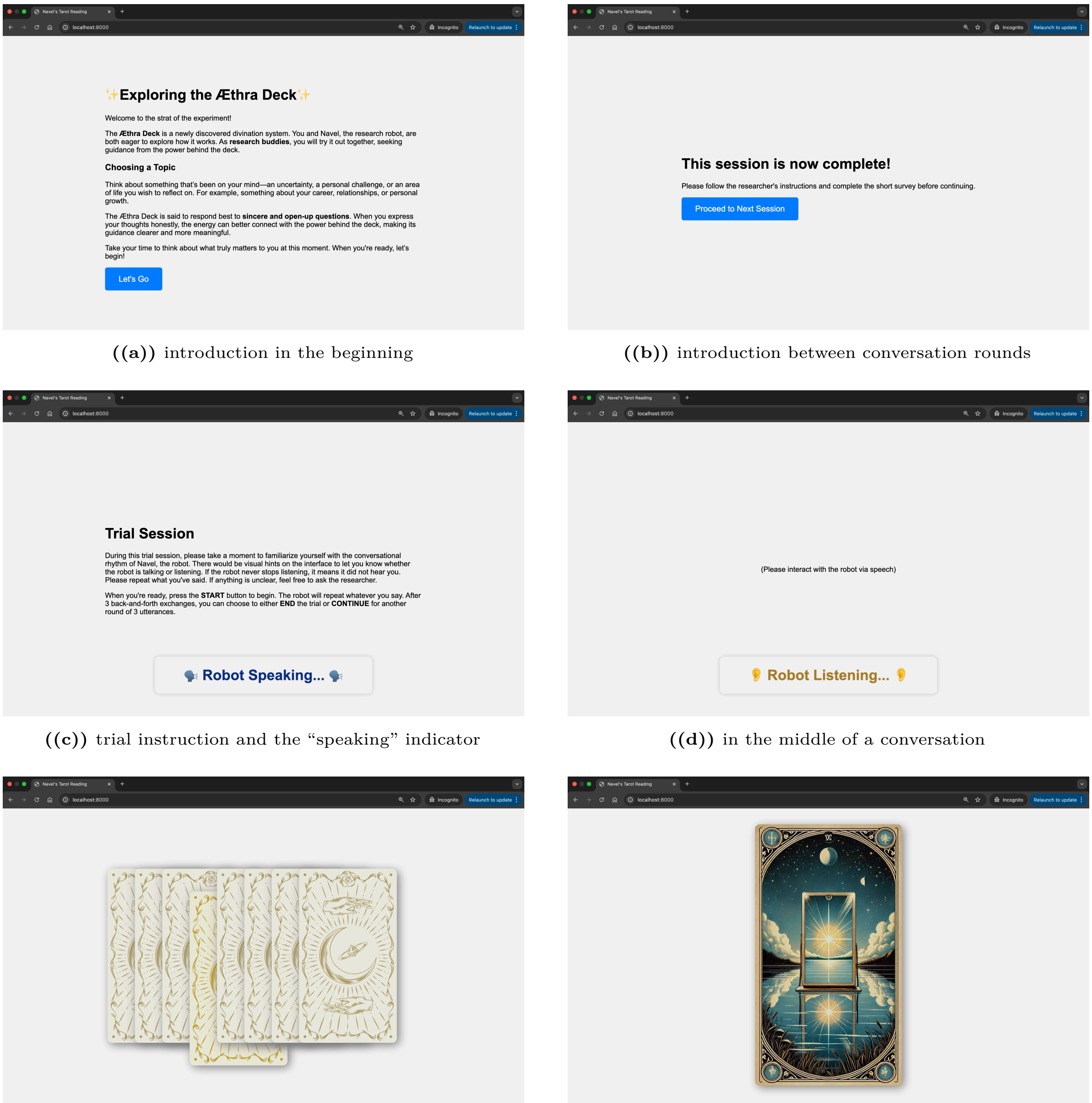
The turn indicator was added following feedback from the internal pilot study. Participants reported difficulty in knowing whether their speech had been detected successfully. In response, a visual indicator was added to continuously display the robot's current state—i.e., whether it is listening or speaking (Figure 3.5 (c)(d)). This helps users avoid speaking when the robot is talking and encourages them to repeat themselves when their speech was not recognized. This small addition nicely improves interaction clarity and user confidence.

Card and Deck

The card selection interface (Figure 3.5(e)) presents eight identical card backs from which the user may choose. A simple hover animation provides visual feedback. Once a card is selected, the corresponding pre-defined card for that round is revealed (Figure 3.5(f)). The card visual remains on the screen for the rest of the round.

[4] https://azure.microsoft.com/en-us/products/ai-services/ai-speech

[5] https://doc.navelrobotics.com/



✨Exploring the Æthra Deck✨

Welcome to the strat of the experiment!

The **Æthra Deck** is a newly discovered divination system. You and Navel, the research robot, are both eager to explore how it works. As **research buddies**, you will try it out together, seeking guidance from the power behind the deck.

Choosing a Topic

Think about something that's been on your mind—an uncertainty, a personal challenge, or an area of life you wish to reflect on. For example, something about your career, relationships, or personal growth.

The Æthra Deck is said to respond best to **sincere and open-up questions**. When you express your thoughts honestly, the energy can better connect with the power behind the deck, making its guidance clearer and more meaningful.

Take your time to think about what truly matters to you at this moment. When you're ready, let's begin!

((**a**)) introduction in the beginning

This session is now complete!

Please follow the researcher's instructions and complete the short survey before continuing.

((**b**)) introduction between conversation rounds

Trial Session

During this trial session, please take a moment to familiarize yourself with the conversational rhythm of Navel, the robot. There would be visual hints on the interface to let you know whether the robot is talking or listening. If the robot never stops listening, it means it did not hear you. Please repeat what you've said. If anything is unclear, feel free to ask the researcher.

When you're ready, press the **START** button to begin. The robot will repeat whatever you say. After 3 back-and-forth exchanges, you can choose to either **END** the trial or **CONTINUE** for another round of 3 utterances.

((**c**)) trial instruction and the "speaking" indicator

(Please interact with the robot via speech)

((**d**)) in the middle of a conversation

((**e**)) the card deck

((**f**)) the drawn card

Figure 3.5: Screenshots of the user interface across different stages of the interaction.

3.5.4. Conversation Control and Synchronization

As the central controller of the conversation flow, the server is connected to the two clients via Socket.IO[6] over the same local network. Socket.IO is preferred over traditional HTTP(s) because it enables real-time, bidirectional communication, ensuring a smoother interaction flow. For example, when a user presses a button on the screen, the server immediately detects the action and notifies the robot to proceed to the next step, eliminating the need for periodic polling. In this architecture, both clients remain stateless, responding only to server requests. Computational tasks are centrally handled by the server.

However, a challenge arose regarding synchronization. Events communicated via socket connections are inherently stateless and unordered. Furthermore, certain operations, such as the robot speaking a sentence or the user providing input via the web UI, require non-negligible time. To ensure that the interaction flow remains deterministic, the server must "wait" for specific signals from the clients before proceeding to the next step. At the same time, the server must remain responsive to other asynchronous events without blocking. To address this, an event queue mechanism was implemented on the server.

3.5.5. Use of Large Language Model

A large language model is used to understand the user's speech and generate tailored utterances. OpenAI GPT-4 was chosen for its simplicity of usage and stable response quality. Unlike many LLM-based chatbots, this system does not refer to the full conversation history when generating responses. Instead, each LLM call only considers the latest user utterance and, if applicable, specific extracted variables. This approach optimizes performance, reduces unnecessary data exchange, and helps preserve the controlled conversational structure.

Information Extraction

Key pieces of information are extracted and stored for later reference, including the user's name, proposed topic, and detailed elaborations. The extracted information is stored as defined variables in runtime memory. Some variables (e.g., user name) are stored across multiple rounds.

The following is an example prompt of extracting the user's name:

```
You are a helpful assistant that extracts information from user input. Please extract the
user's name. Only return the name itself. If the name cannot be extracted, return 'buddy'.
```

Response Generation

As presented in Section 3.4.1, the robot refers to the user's speech at predefined points in the conversation. The LLM is also used to generate such responses. For example, when asking the user to elaborate on their topic the following prompt is used:

```
You are a helpful assistant that guides the user to talk about their personal issues. The
user inputs a topic. You ask for more details about that topic. Only return the question
itself in one sentence.
```

In some cases, the robot refers to both the immediate user response and earlier extracted information. For instance, a tradeoff based on previous context is summarized with the following prompt, where `<topic>` and `<details>` are previously extracted variables:

```
You are a helpful assistant that summarizes information from user input. You two are
discussing the user's personal issue <topic>, with details: <details>. The user will talk
about possible tradeoffs. Please summarize the information in one sentence, like 'You have
mentioned that... and the tradeoff...'.
```

The complete set of prompt used to generate responses can be seen in Appendix B.

[6] https://socket.io/

Social Attitude

To ensure that the robot's tone matches the assigned condition (social or baseline), different background instructions are set for the LLM prompts. In the social condition, it is:

```
You are an empathetic and cheerful agent who is replying to the user's opinion. You two are
discussing about the content of a divination card. You always encourage and acknowledge
the users opinion or plan. Sometimes, you also show your own emotions.
```

As for the baseline condition, it is instead:

```
You are an cool and professional agent who is replying to the user's opinion. You two are
discussing about the content of a divination card. You never call the user by their name.
You never talk about feelings or emotions, and always act neutral instead of giving own
opinion.
```

3.6. Internal Pilot Study

To ensure that the designed and implemented system is ready for the user study, an internal pilot study was conducted. Two participants were recruited to experience the full interaction procedure under the social condition. Several improvements were made based on observations and feedback.

- **Turn Indicator**: A real-time turn indicator was added to the user interface, showing whether the robot was currently speaking or listening. This aimed to reduce user uncertainty about the system's state during interaction. (See Section 3.5.3)
- **Trial Round**: The trial interaction was enriched to better simulate the main interaction experience. Initially, the robot simply repeated the user's utterances to test input/output functionality. In the updated version, a random small talk question is appended, allowing users to practice real-time thinking and responding. (See Section 4.1)
- **Instruction Wording**: Instructional texts and survey item descriptions were revised for improved clarity and conciseness. For instance, terms such as "social robot" and "conversational agent" are now explicitly defined within the questionnaire. (See Appendix C)
- **Conversation Length**: Utterances were slightly shortened to better balance the speaking time between the two parties. This adjustment also maintained each round's duration within approximately 4-5 minutes.

The internal pilot confirmed the overall feasibility of the system and guided minor refinements before launching the formal user study.

4 User Study

In the previous chapter, the design and development of the card-based interaction were thoroughly described. This chapter presents the main user study in which the interaction is applied. This controlled experiment serves as the core methodological tool to address the primary research question: *How does multidimensional trust in a social robot develop over time?*

4.1. Study Design

This study adopts a 2×3 mixed design. The between-subject factor is the robot's social attitude (*social* vs. *baseline*). The within-subject factor is time point, with repeated measurements taken after each of the three conversation rounds (t_1, t_2, t_3). The dependent variables include cognitive and affective trust, certainty in trust assessment and subjective topic intimacy.

4.1.1. Conditions

The robot's behavior varies depending on the assigned social attitude condition, which is manipulated through both verbal and non-verbal channels, as inspired by the literature (see Section 2.4.1). The *verbal* manipulations are based on the conversation scripts and LLM prompts described in Section 3.4, ensuring the two conditions differ only in social expressiveness while keeping semantic content and flow comparable. In the *social* condition, the robot speaks in a friendly and empathetic tone and refers to previous rounds to create a sense of familiarity. In contrast, the *baseline* robot maintains a neutral and professional language, treating each round as an isolated encounter without memories of shared history.

On the *non-verbal* level, facial expression is the primary cue. By default, the Navel robot exhibits subtle expressions like blinking, raising eyebrows, or smiling while speaking. In the *social* condition, additional expression tags, such as smiling, neutral, or curious, are inserted into the script to be triggered at specific moments. These are listed in Appendix B. In the *baseline* condition, all facial expressions are disabled, including the default animations.

Originally, we considered using gaze behavior as an additional social cue, given that is a widely acknowledged social feature in for both interpersonal and HRI fields [12, 60, 6]. However, due to the instability of the face detection system, which often caused the robot's head to jitter while attempting to maintain eye contact, this feature was removed. Participants in the pilot study described the motion as unnatural and distracting.

4.2. Participants

A total of 45 participants were recruited through personal networks and flyers distributed on campus. Due to technical issues encountered during the study, data from 5 participants were excluded. The final sample consisted of 40 participants (19 female, 20 male, 1 non-binary), most of whom were university students or recent graduates, aged between 21 to 35.

All participants reported normal or corrected-to-normal hearing and vision, and demonstrated sufficient English proficiency to comfortably engage in spoken conversations. Based on gender and age, participants were evenly distributed across the two experimental groups to ensure demographic balance. To thank participants for their involvement, light refreshments (e.g., snacks) were provided at the end of the study. No monetary compensation was offered.

4.3. Measurements

To investigate how participants' trust in the robot develops over time, self-reported metrics are collected for quantitative analysis. Additionally, open-ended questions are included to gain deeper insights into participants' underlying thoughts and reasoning. The different measurements are described in more detail below. Complete set of survey questions and the corresponding instructions are presented in Appendix C.

4.3.1. Demographics and Control Variables

Demographics

In addition to basic demographic information such as gender and age, participants are asked about their prior experience with social robots and conversational agents. They report the frequency of interaction and the typical contexts in which such interactions occur. These responses are used for descriptive purposes and to explore potential individual differences in trust perception.

Paranormal Beliefs

As discussed in 3.2.1, participants' beliefs in supernatural phenomena are considered a possible confounding variable. To account for this, the Precognition subscale of the Revised Paranormal Belief Scale (RPBS) [59] is administered. This subscale was selected for its relevance to the study context, unlike other subscales such as Witchcraft (e.g., *"Black magic really exists"*) or Psychokinesis (e.g., *"A person's thoughts can influence the movement of a physical object"*).

To better align with the task framing, one item *"Astrology is a way to accurately predict the future"* was rephrased as *"Tarot is a way to accurately reveal guidance about the future."* The average score is calculated and used as a control variable to examine potential moderation effects on trust formation.

4.3.2. Post-Interaction Survey

After each of the three interaction rounds, participants complete a short survey reflecting on their experience with the conversation, the card deck, and the robot.

Originally, a pre-interaction measurement was planned as an additional data point. However, most items in the affective/cognitive trust questionnaires were not applicable before any interaction occurred (e.g., *"Interacting with Navel, I have no reservation about acting on its advice"*). Due to concerns about facial validity, this pre-interaction measurement was removed from the procedure.

Subjective Topic Intimacy

To capture how personally meaningful each conversation was, participants are asked to rate how intimate the topic they discussed was (0–100%). This serves as an indirect indicator of affective trust and complements the lexical-based disclosure intimacy metrics.

Impression and Trust in the Card

Participants evaluate their impression of the card deck using two items (1–7 Likert scale):

1. *"I believe the Æthra deck could offer meaningful guidance on my concerns."*
2. *"The symbolism of the Æthra deck made sense on its own."*

These questions help isolate participants' trust in the robot from their trust in the card system, as discussed in 3.2.1. The responses also allow for exploratory analysis of whether perceptions of the deck are associated with robot trust ratings.

Trust in the Robot

The participants' trust in the robot is the central focus of this study. This measurement follows the approach of Anzabi and Umamero [6], utilizing an adapted version of the cognitive-affective trust questionnaire. The original scale, developed by McAllister [42] and Johnson and Grayson [24], has been widely applied in both interpersonal and HRI contexts (e.g., [6, 5]). Consistent with Anzabi and Umamero's adaptation, terms such as "this person" were replaced with "Navel", and "work" with "interact" to match the study context.

In line with RQ(c), participants are also asked to rate their certainty in each trust assessment (0–100%) immediately after completing the Likert-scale items for either cognitive or affective trust. An optional open-ended question follows, inviting them to explain the reasons behind their ratings. These responses allow investigation into how dynamic-learned trust becomes more consolidated over time.

Alternative trust measures developed specifically for human-robot interaction, such as MDMT v2 [40] and the Trust Perception Scale-HRI [56], were initially considered. However, the cognitive-affective trust scale was ultimately selected due to its relevance to social interaction and its clear distinction between cognitive and affective dimensions.

4.3.3. Post-Study Open Questions

At the end of the experiment, participants answered several open-ended questions to retrospect on their trust perception and the overall interaction across multiple rounds. The instructions guide participants to describe how their feelings and impressions were during each conversation round, and how they may have changed over time. These questions aim to uncover the reasoning behind the quantitative ratings and to provide additional insight into the process of trust development. Table 4.1 summarizes the questions along with their corresponding research purposes.

ID	Question	Purpose
OP1	*How did your impression of* ***Navel's understanding of you*** *(i.e. if Navel got to know you better) change over the course of the sessions, if at all?*	Manipulation check: if memory and personalization are perceived.
OP2	*How did your* ***affective trust*** *in Navel change as the sessions progress, if at all?* *Affective trust refers to the belief that Navel genuinely cares about your well-being, and is emotionally connected and close to you. In other words, did you feel more or less comfortable to emotionally rely on Navel over time?*	To understand participants' reasoning about affective trust assessment. (RQ(a))
OP3	*How did your* ***cognitive trust*** *in Navel change as the sessions progress, if at all?* *Cognitive trust refers to your confidence in the Navel's competence, reliability, and professionalism. In other words, did your perception of Navel's ability and reliability improve or decline over the sessions?*	To understand participants' reasoning about cognitive trust assessment. (RQ(b))
OP4	*[Optional] Is there any additional feedback for the whole experiment?*	Additional feedback

Table 4.1: Post-study open questions and their research purpose

4.4. Procedure

As illustrated in Figure 4.1, the entire study took approximately 30-45 minutes. Each participant arrived individually and was welcomed into a quiet, isolated room. The researcher provided a brief overview of the study and invited the participant to read and sign the informed consent form, and to complete the pre-experiment questionnaires.

Next, the participant engaged in a brief trial round designed to familiarize them with (1) the conversational rhythm with the robot, and (2) the user interface on the screen. The trial consisted of three short dialogue rounds. In each round, the robot initiated a small-talk prompt (e.g., *"How was your*

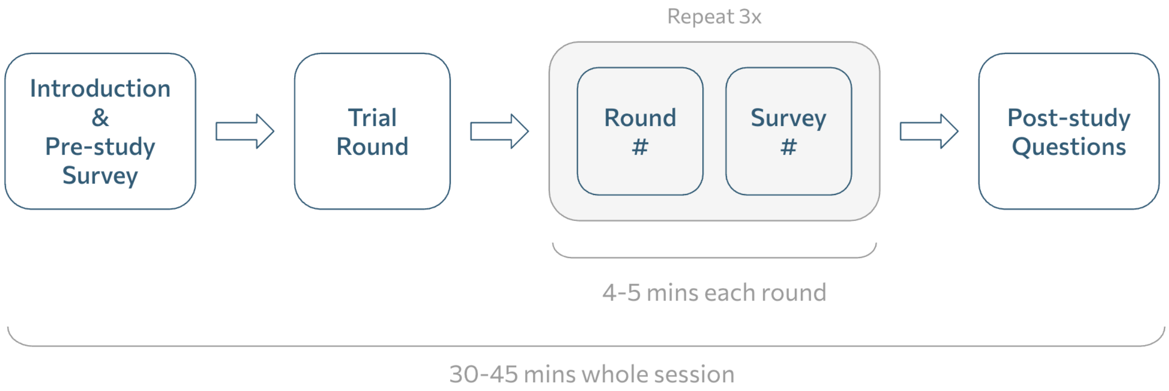


Figure 4.1: The overall study procedure

day?"). After the participant responded, the robot repeated their response to confirm that their speech had been recognized. If the participant felt they needed more practice, they could opt to repeat an additional set of three rounds.

Once the participant indicated they were ready, the main experimental began. In each of the three interaction rounds, they were invited to share a personal concern, draw a symbolic card, and reflect on its meaning together with the robot. Each interaction lasted approximately 4-5 minutes. The detailed structure of each interaction round is described in 3.4.1.

Since the conversation content could involve sensitive or personal topics, the researcher did not stay in the room during the interactions. The room was left private for the participant and the robot to ensure a comfortable setting. However, the researcher continuously monitored the process through a backend terminal to ensure that any technical issues could be addressed promptly. After each conversation round, the researcher reentered the room to provide the post-interaction survey and check in with the participant.

After all three interaction rounds were completed, the participant answered a final post-study questionnaire. Then, the researcher explained that the symbolic cards were not magical or tailored, in order to ethically clarify the nature of the task. Participants were given the opportunity to ask questions or share feedback. Finally, they were thanked for their participation with a small snack and guided to leave.

4.5. Ethical Considerations

As the experiment was designed to be personal and intimate, it was crucial to ensure that participants remained unidentifiable and felt safe disclosing personal experiences or concerns. The study protocol was developed in accordance with the General Data Protection Regulation (GDPR)[1] and the Human Research Ethics Committee (HREC) guidelines of the Delft University of Technology[2]. The protocol received ethical approval from the TU Delft HREC with ID 5502.

Although the interactions were speech-based, only real-time transcripts generated by the Azure Speech-to-Text service were collected. No audio recordings were stored at any point. To further protect participant identity, all transcripts were pseudonymized. In the social condition, the robot referred to the participant by their name to foster familiarity. However, the name was stored only in runtime memory during the experiment and was automatically replaced with the string "[name]" in all logged transcripts. Furthermore, the researcher manually reviewed all transcripts to identify and pseudonymize potentially identifying information before archiving the dataset. For instance, a statement such as "I am Sam, I have autism" would be logged as "I am [name], I have [health condition]".

Informed consent was obtained from all participants prior to the experiment (See Appendix C.1).

[1] https://gdpr-info.eu/
[2] https://www.tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/human-research-ethics

The consent form explicitly offered participants the option to opt out of public data archiving (i.e., 4TU.ResearchData) while still participating in the study. One participant chose to do so. At the end of the study, all participants were given the opportunity to review their transcripts and request removal or further pseudonymization of any content. However, none of the participants made such requests.

Finally, given that religious and philosophical beliefs are classified as a "special category of personal data" under GDPR, we discussed the collection of paranormal belief in advance with TU Delft's privacy team [3]. Since the study only assessed general attitudes along the skeptical–spiritual spectrum without referencing specific religions, the privacy team confirmed that no additional data protection measures or ethical applications were required.

[3] https://www.tudelft.nl/data-protection/privacy

5

Results

This chapter presents the results of the user study. We begin by examining the group balance in participants' demographic characteristics (Section 5.1). Then, we present a quantitative analysis on trust development (Section 5.2) and certainty in trust assessment (Section 5.3), followed by an exploratory correlation analysis on contextual factors (Section 5.4). Finally, we show the qualitative results (Section 5.5).

5.1. Participants

As described in Section 4.2, participants of different genders and age groups were balanced across the experimental conditions. Figure 5.1 illustrates the distributions of gender, age, and prior experience with social robots or conversational agents, which appear similar across groups. To confirm this, we conducted statistical tests.

A Chi-squared test was used to examine gender distribution across groups, revealing no significant difference ($\chi^2 = 1.674$, $p = .431$). Kruskal–Wallis rank-sum tests were conducted for the ordinal variables. No significant differences were found for age group ($H = .008$, $p = .928$), or prior experience with social robots or conversational agents, ($H = .002$, $p = .967$). These results suggest that demographic characteristics were well balanced across the two social attitude conditions.

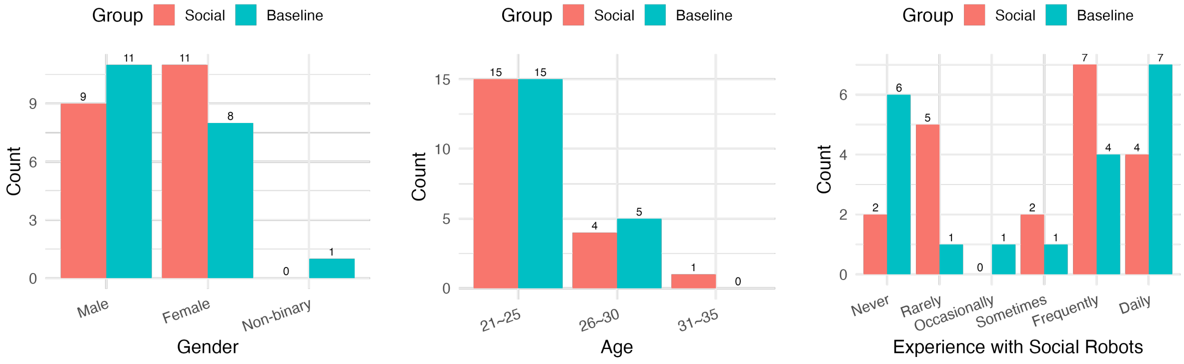


Figure 5.1: The distribution of participants' demographics data

5.2. Trust Development

This section presents the results for participants' cognitive and affective trust in Navel. Figure 5.2 illustrates the development of the two trust dimensions across time in each experimental group. The corresponding means and standard deviations are reported in Table 5.1. Table 5.2 summarizes the results of mean-based statistical tests.

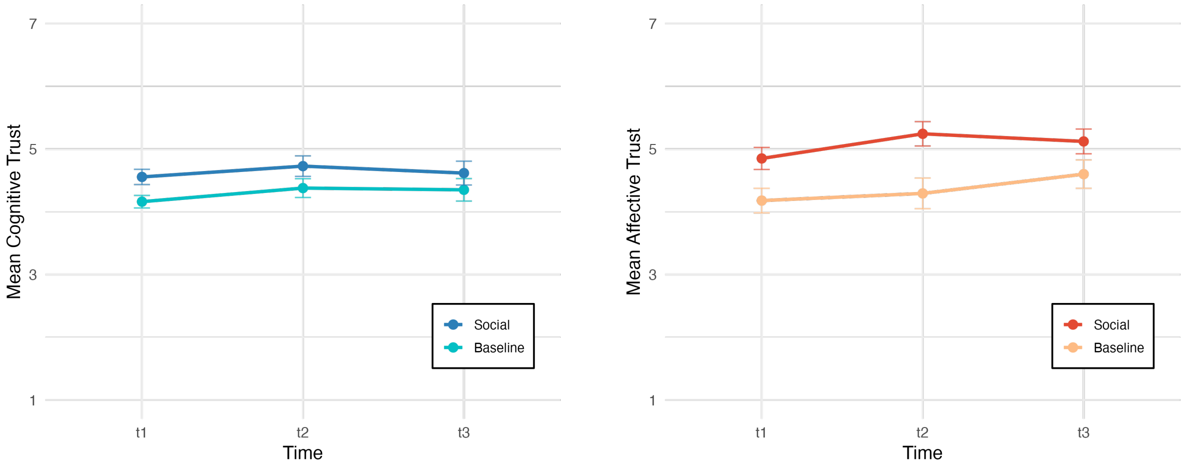


Figure 5.2: The development of cognitive and affective trust under different social conditions across time

Trust Type	Social Attitude	t_1	t_2	t_3	Overall
Cognitive Trust	Social	4.56 (0.544)	4.73 (0.730)	4.62 (0.849)	4.63 (0.710)
	Baseline	4.16 (0.447)	4.38 (0.669)	4.35 (0.802)	4.30 (0.652)
	Overall	4.36 (0.530)	4.55 (0.714)	4.48 (0.826)	4.46 (0.700)
Affective Trust	Social	4.85 (0.784)	5.24 (0.868)	5.12 (0.882)	5.07 (0.848)
	Baseline	4.18 (0.878)	4.29 (1.091)	4.60 (1.010)	4.36 (0.996)
	Overall	4.51 (0.889)	4.77 (1.086)	4.86 (0.972)	4.71 (0.989)

Table 5.1: The mean (and SD) of cognitive and affective trust across time.

DV	Test	Variable	Statistic	p	Effect Size
Cognitive Trust	Mixed ANOVA	Social attitude × Time	$F = 0.265$	.768	$ges = .002$ (N)
		Time	$F = 2.447$	$.093^{\dagger}$	$ges = .014$ (S)
		Social attitude	$F = 3.089$	$.087^{\dagger}$	$ges = .059$ (S)
Affective Trust	Mixed ANOVA	Social attitude × Time	$F = 1.814$	.170	$ges = .010$ (N)
		Time	$F = 4.933$	$.010^{*}$	$ges = .026$ (S)
		Social attitude	$F = 7.491$	$.009^{*}$	$ges = .136$ (M)
	(Post-hoc) pairwise t-tests	$t_1 < t_2$	$t = 2.18$	.107	$d = .249$ (S)
		$t_2 < t_3$	$t = 0.78$	1.000	$d = .089$ (N)
		$t_1 < t_3$	$t = 3.14$	$.010^{**}$	$d = .370$ (S)
Cognitive Certainty	Mixed ANOVA	Social attitude × Time	$F = 1.183$	.312	$ges = .010$ (N)
		Time	$F = 11.967$	$< .001^{***}$	$ges = .090$ (M)
		Social attitude	$F = 2.837$	$.100^{\dagger}$	$ges = .049$ (S)
	(Post-hoc) pairwise t-tests	$t_1 < t_2$	$t = 3.66$	$.002^{**}$	$d = .543$ (M)
		$t_2 < t_3$	$t = 0.61$	1.000	$d = .089$ (N)
		$t_1 < t_3$	$t = 4.33$	$< .001^{***}$	$d = .637$ (M)
Affective Certainty	Wald-type test	Social attitude × Time	$W = 3.832$	.147	—
		Time	$W = 13.755$	$.001^{**}$	—
		Social attitude	$W = 5.967$	$.015^{*}$	—
	(Post-hoc) Wilcoxon signed-rank	$t_1 < t_2$	$V = 113$	$.002^{**}$	—
		$t_2 < t_3$	$V = 316$	1.000	—
		$t_1 < t_3$	$V = 184$	$.035^{*}$	—

Table 5.2: Summary of mean-based statistical tests on trust and certainty-related dependent variables. Post-hoc comparisons were Bonferroni-corrected. Interpretation of *ges* (generalized eta-squared) are: $negligible(N) < .01$, $small(S) = .02$, $medium(M) = .13$, $large(L) = .26$; Interpretation of d (Cohen's d) are: $negligible(N) < .20$, $small(S) = 0.2 \sim 0.3$, $medium(M) \approx 0.5$, $large(L) > 0.8$.

5.2.1. Cognitive Trust

We began by checking the assumptions for parametric analysis. A Shapiro-Wilk test confirmed normality in both groups (social: $p = .718$, baseline: $p = .055$), and Levene's test indicated homogeneity of variances ($F = .842$, $p = .361$). A mixed-design ANOVA was conducted to examine the effects of social attitude (between-subjects) and time (within-subjects) on cognitive trust.

The results revealed no significant interaction effect ($F = .265$, $p = .768$, $ges = .002$). The main effect of time was marginally significant ($F = 2.447$, $p = .093$), but the small effect size ($ges = .014$) suggests limited practical relevance. The main effect of social attitude was also marginally significant ($F = 3.089$, $p = .087$) with a medium effect size ($ges = .059$), indicating a trend where participants in the social condition reported higher cognitive trust ($M = 4.63$, $SD = .710$) than those in the baseline condition ($M = 4.30$, $SD = .652$).

A primary focus of this research is to understand how trust may develop over time. While the ANOVA test identifies whether there are significant differences in trust scores across time, it does not capture the developmental trajectory. Therefore, we additionally conducted a linear mixed-effects model (LMM) analysis to examine the evolution of trust across time. The LMM also allows us to explore potential confounding variables.

Two models were fitted:

$$M_{CT0}\text{: cognitive trust} \sim \text{social attitude} * \text{time} + (1 \mid \text{id}) \quad \text{(selected)}$$

$$M_{CT1}\text{: cognitive trust} \sim \text{social attitude} * \text{time} + \text{paranormal} + (1 \mid \text{id})$$

Both models included social attitude, time, and their interaction as fixed effects, with participant ID as a random intercept. Model M_{CT1} additionally included participants' paranormal belief scores as a potential confounding variable. However, its model fit ($AIC = 207.85$) was slightly worse than that of M_{CT0} ($AIC = 207.05$), so M_{CT0} was retained as the final model. The fixed effect estimates of M_{CT0} are shown in Table 5.3.

Fixed Effect	Estimate	SE	df	t	p	
(Intercept)	4.30	0.14	38.00	31.69	< .001	***
social attitude (Social)	0.34	0.19	38.00	1.76	.0869	.
time (Linear)	0.13	0.09	76.00	1.50	.1379	
time (Quadratic)	-0.10	0.09	76.00	-1.12	.2661	
social attitude × time (Linear)	-0.09	0.13	76.00	-0.72	.4754	
social attitude × time (Quadratic)	-0.02	0.13	76.00	-0.13	.9000	

Table 5.3: Fixed effect estimates from the LMM M_{CT0} predicting *cognitive trust*. Participant ID was included as a random intercept. (Random intercept $SD = 0.561$; residual $SD = 0.398$)

In line with the ANOVA results, the LMM indicates that participants in the social group had marginally significantly higher cognitive trust than those in the baseline group. However, no effects of time, either linear or quadratic, were found, suggesting that cognitive trust remained stable over time.

5.2.2. Affective Trust

Prior to analyzing affective trust, normality and homogeneity assumptions were assessed. Levene's test confirmed equal variances across groups ($F = .473$, $p = .493$). The Shapiro-Wilk test on raw scores indicated that the data was normally distributed in the social group ($p = .783$), but not in the baseline group ($p = .004^{**}$). However, a Shapiro-Wilk test on the residuals of the ANOVA model confirmed that the residuals were approximately normally distributed ($p = .535$), satisfying the assumption for parametric testing. Given that mixed-design ANOVA is generally robust to minor violations of normality, we proceeded with the analysis.

The ANOVA showed no significant interaction effect ($F = 1.814$, $p = .170$, $ges = .010$), but revealed a significant main effect of time ($F = 3.511$, $p = .010^{**}$) with a small to medium effect size ($ges = .026$). To further investigate the main effect of time, pairwise t-tests with Bonferroni correction were conducted.

Results indicated that affective trust at t_3 ($M = 4.86$, $SD = .972$) was significantly higher than at t_1 ($M = 4.51$, $SD = .889$; $p = .010^{**}$, $d = .370$). No significant differences were found between t_1 and t_2 ($p = .107$, $d = .249$) or t_2 and t_3 ($p = 1.000$, $d = .089$).

The main effect of social attitude is significant ($F = 7.491$, $p = .009^{**}$) with a medium to large effect size ($ges = .136$). Participants in the social group reported higher affective trust ($M = 5.07$, $SD = .848$) than those in the baseline group ($M = 4.36$, $SD = .996$).

Similarly, two linear mixed models were constructed to examine affective trust, with social attitude, time, and their interaction as fixed effects and participant ID as a random intercept:

$$M_{AT0}\text{: affective trust} \sim \text{social attitude} * \text{time} + (1 \mid \text{id}) \qquad \text{(selected)}$$

$$M_{AT1}\text{: affective trust} \sim \text{social attitude} * \text{time} + \text{paranormal} + (1 \mid \text{id})$$

Model M_{AT0} was ultimately selected, as it showed slightly better fit ($AIC = 271.43$) than M_{AT1} ($AIC = 273.31$), suggesting that individual belief in the paranormal did not contribute meaningfully to the model. Table 5.4 presents the fixed effect estimates of M_{AT0}.

Fixed Effect	Estimate	SE	df	t	p	
(Intercept)	4.36	0.18	38.00	23.61	< .001	***
social attitude (Social)	0.71	0.26	38.00	2.74	.0094	**
time (Linear)	0.30	0.11	76.00	2.61	.0109	*
time (Quadratic)	0.08	0.11	76.00	0.69	.4926	
social attitude × time (Linear)	-0.11	0.16	76.00	-0.66	.5132	
social attitude × time (Quadratic)	-0.29	0.16	76.00	-1.79	.0778	.

Table 5.4: Fixed effect estimates from the LMM M_{AT0} predicting *affective trust*. Participant ID was included as a random intercept. (Random intercept $SD = 0.771$; residual $SD = 0.511$)

In addition to the significant main effect of social attitude consistent with the ANOVA results, the LMM revealed a significant linear trend across time points. This trend indicates that affective trust increase as participants progressed through time. Furthermore, the quadratic interaction between social attitude and time was marginally significant, suggesting that the two groups exhibited distinct developmental trajectories of affective trust. In the social group, trust increased sharply from t_1 to t_2, followed by a slight decline at t_3. In contrast, the baseline group showed relatively stable trust between t_1 and t_2, with a subsequent increase at t_3. These findings suggest that affective trust might develop differently depending on the robot's social behavior.

5.2.3. Summary

Taken together, the results reveal distinct developmental patterns for different trust dimensions. While affective trust demonstrated an increase over time, cognitive trust remained stable. Additionally, the marginally significant quadratic interaction between social attitude and time suggests that affective trust development followed distinct trajectories depending on the robot's social behavior.

The robot's social attitude tend to have positive influences on both cognitive (marginally) and affective (significantly) trust, with participants in the social condition generally reporting higher levels of trust. Paranormal belief, on the other hand, did not meaningfully contribute to predicting trust levels in either dimension.

5.3. Certainty in Trust Assessment

We measured participants' certainty in both their cognitive and affective trust judgments. The means and standard deviations are reported in Table 5.5 and illustrated in Figure 5.3. Mean-based statistical tests are summarized in Table 5.2. As stated in **RQ(c)**, the primary aim was to examine whether certainty evolved over time. Nonetheless, we still explored potential effects of social attitude and interaction.

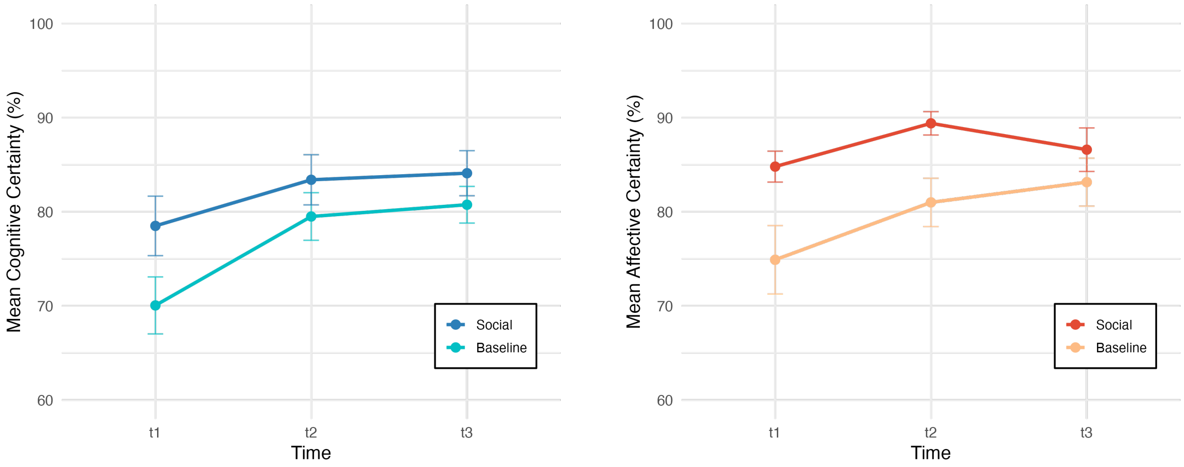


Figure 5.3: Participants' certainty in cognitive and affective trust assessment over time

Trust Type	Social Attitude	t_1	t_2	t_3	Overall
Certainty in Cognitive Trust	Social	78.50 (14.137)	83.40 (11.927)	84.10 (10.701)	82.00 (12.384)
	Baseline	70.05 (13.547)	79.50 (11.316)	80.75 (8.711)	76.77 (12.165)
	Overall	74.28 (14.320)	81.45 (11.644)	82.42 (9.779)	79.38 (12.502)
Certainty in Affective Trust	Social	84.80 (7.346)	89.40 (5.548)	86.60 (10.333)	86.93 (8.082)
	Baseline	74.90 (16.274)	81.00 (11.475)	83.15 (11.389)	79.68 (13.486)
	Overall	79.85 (13.433)	85.20 (9.861)	84.88 (10.875)	83.31 (11.654)

Table 5.5: The mean (and SD) of certainty in cognitive and affective trust assessment across time.

5.3.1. Certainty in Cognitive Trust Assessment

Levene's test confirmed the homogeneous assumption for mixed ANOVA ($F = .403$, $p = .527$). The Shapiro-Wilk test revealed normality in the social group ($p = .200$), but not in the baseline ($p = .001^{***}$). Once again, given the robust of mixed-design ANOVA and the normality in the residuals ($p = .735$) shown by another Shapiro-Wilk test, we proceeded with the parametric analysis.

The mixed ANOVA test shows no significant interaction effect ($F = 1.183$, $p = .312$, $ges = .010$). A significant main effect of time was found ($F = 11.967$, $p = .001^{***}$) with a medium to large effect size ($ges = .090$). Post-hoc pairwise pairwise t-tests with Bonferroni correction revealed that certainty significantly increased: (1) from t_1 to t_2 ($t = 3.66$, $p = .002^{**}$) with medium effect size ($d = .543$), and (2) from t_1 to t_3 ($t = 4.33$, $p < .001^{***}$) with medium effect size ($d = .637$). No significant difference was found between t_2 and t_3 ($t = .605$, $p = .549$, $d = .089$), indicating that certainty plateaued after t_2.

The main effect of social attitude is just marginal ($F = 2.837$, $p = .100$). However, the small to medium effect size ($ges = .049$) suggests that such effect could still be meaningful and that participants in the social group tend to report higher confidence in cognitive trust judgment.

Initially, we only assumed time to affect participants' certainty in trust judgment. However, as the ANOVA test has shown a potential difference between groups, we took social attitude into account when fitting linear mixed-effects models. A total of four models were fitted:

$$M_{CC0}\text{: certainty in cognitive trust} \sim \text{time} + (1 \mid \text{id})$$

$$M_{CC1}\text{: certainty in cognitive trust} \sim \text{time} + \text{paranormal} + (1 \mid \text{id}) \qquad \text{(selected)}$$

$$M_{CC2}\text{: certainty in cognitive trust} \sim \text{social attitude} * \text{time} + (1 \mid \text{id})$$

$$M_{CC3}\text{: certainty in cognitive trust} \sim \text{social attitude} * \text{time} + \text{paranormal} + (1 \mid \text{id})$$

With the least AIC (M_{CC0}: 911.89, M_{CC1}: 911.40, M_{CC2}: 912.56, M_{CC3}: 911.45), M_{CC1} was eventually selected for further analysis. As shown in Table 5.6, the model revealed a significant linear effect of time ($p < .001^{***}$), indicating that participants' certainty in their cognitive trust judgments increased over time. Additionally, a marginal quadratic effect was observed ($p = .053$), suggesting that the rate of increase might slow down as time progressed. The effect of paranormal belief was not significant ($p = .127$), indicating that individual belief levels did not predict certainty ratings.

Fixed Effect	Estimate	SE	df	t	p	
(Intercept)	73.45	4.11	38.00	17.88	< .001	***
time (Linear)	5.76	1.29	78.00	4.47	< .001	***
time (Quadratic)	-2.53	1.29	78.00	-1.96	.0532	.
paranormal belief	1.84	1.18	38.00	1.56	.1267	

Table 5.6: Fixed effect estimates from the LMM M_{CC1} predicting *certainty in cognitive trust.* Participant ID was included as a random intercept. (Random intercept SD = 8.68; residual SD = 8.16)

5.3.2. Certainty in Affective Trust Assessment

The data passed neither normality nor homogeneity check. Thus, the non-parametric Wald-type test was used. No significant interaction effect ($W = 3.832$, $p = .241$) was found.

A significant main effect of time was observed ($W = 13.755$, $p = .001^{***}$). Post-hoc Wilcoxon signed-rank tests with Bonferroni correction revealed a significant increase in certainty from t_1 to t_2 ($V = 113$, $p = .002^{**}$) and from t_1 to t_3 ($V = 184$, $p = .035^{*}$). Difference between t_2 and t_3 ($V = 316$, $p = 1.000$) was not significant, suggesting that the average certainty increased early and then stabilized.

The main effect of social attitude was also significant ($W = 5.967$, $p = .015^{*}$), showing that participants interacting with a socialized robot were more confident about their assessment in affective trust.

Similar to cognitive certainty, four linear mixed-effects models were considered:

$$M_{AC0}\text{: certainty in affective trust} \sim \text{time} + (1 \mid \text{id})$$

$$M_{AC1}\text{: certainty in affective trust} \sim \text{time} + \text{paranormal} + (1 \mid \text{id})$$

$$M_{AC2}\text{: certainty in affective trust} \sim \text{social attitude} * \text{time} + (1 \mid \text{id}) \qquad \text{(selected)}$$

$$M_{AC3}\text{: certainty in affective trust} \sim \text{social attitude} * \text{time} + \text{paranormal} + (1 \mid \text{id})$$

M_{AC2} was selected with the smallest AIC (M_{AC0}: 898.49, M_{AC1}: 900.09, M_{AC2}: 894.13, M_{AC3}: 895.40) The estimated fixed effects are reported in Table 5.7. Consistent with the Wald-type test, social attitude emerged as a significant indicator ($p = .015^{*}$). Moreover, the model revealed a highly significant linear effect of time ($p < .001^{***}$), suggesting a strong increase in affective certainty across repeated interactions. Notably, the interaction between social attitude and the linear term of time was marginally significant ($p = .060$), suggesting that the increment of affective certainty may differ between groups.

Fixed Effect	Estimate	SE	df	t	p	
(Intercept)	79.68	2.02	38.00	39.48	< .001	***
social attitude (Social)	7.25	2.85	38.00	2.54	.0153	*
time (Linear)	5.83	1.69	76.00	3.45	.0009	***
time (Quadratic)	-1.61	1.69	76.00	-0.95	.3436	
social attitude × time (Linear)	-4.56	2.39	76.00	-1.91	.0604	.
social attitude × time (Quadratic)	-1.41	2.39	76.00	-0.59	.5578	

Table 5.7: Fixed effect estimates from the LMM M_{AC2} predicting *certainty in affective trust.* Participant ID was included as a random intercept. (Random intercept SD = 7.90; residual SD = 7.57)

5.3.3. Summary

The results revealed a clear temporal trend in participants' certainty about their trust judgments across both trust types, with certainty increasing linearly over time. Mean-based group comparisons (i.e., ANOVA and Wald-type tests) suggested a positive influence of social attitude. However, this effect reached significance only for affective trust, while it remained marginal for cognitive trust.

Linear mixed-effects models further revealed more nuanced developmental patterns. For cognitive trust, a marginal quadratic effect of time suggests a rapid initial increase that quickly plateaued. In contrast, for affective trust, a marginal group-by-time interaction implies that certainty under the baseline condition may have increased more quickly.

5.4. Correlation Analysis

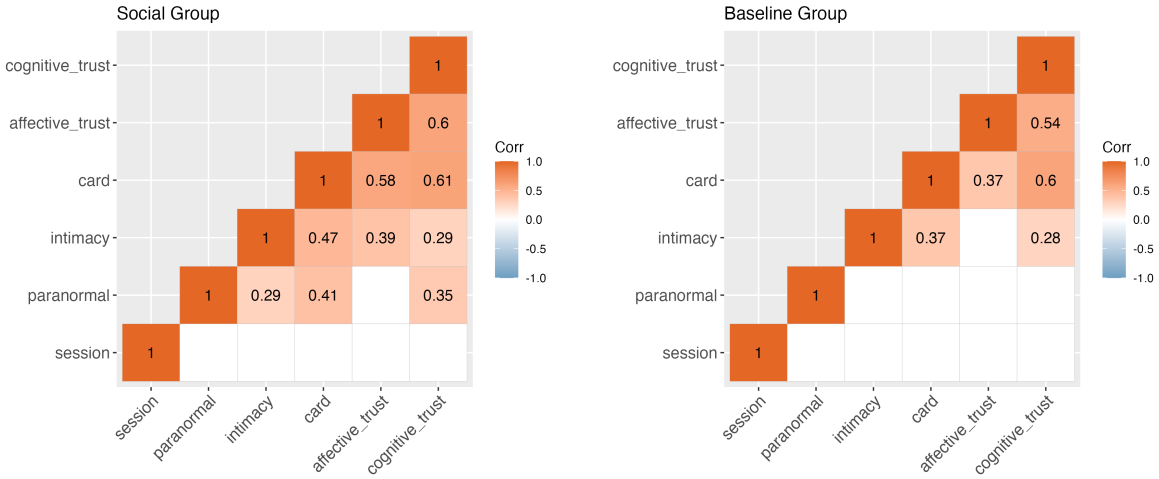


Figure 5.4: The Pearson's correlation value among variables under the social and the baseline condition. Only correlations with significance ($p < .05$) are shown.

Pearson correlation coefficients were computed to examine the relationships among trust-related variables within each experimental group. The variables included affective trust, cognitive trust, topic intimacy, paranormal belief, trust in the card, and session number. For each variable pair, individual correlation tests were performed. Only statistically significant correlations are visualized in Figure 5.4.

Cognitive and affective trust in the robot were strongly correlated in both groups (Social: $r = .60$, Baseline: $r = .54$). Participants' trust in the card also showed a strong association with cognitive trust across conditions (Social: $r = .61$, Baseline: $r = .60$). Yet, its correlation with affective trust was stronger in the social group ($r = .58$) than in the baseline group ($r = .37$).

Under both condition, topic intimacy is moderate correlated with trust in card (Social: $r = .47$, Baseline: $r = .37$) and slightly correlated with cognitive trust (Social: $r = .29$, Baseline: $r = .28$). The correlation with affective trust is significant only in the social group ($r = .39$).

Notably, in the social group, individuals' paranormal belief have small to moderate correlations with cognitive trust ($r = .35$), trust in the card ($r = .41$), and topic intimacy ($r = .29$). However, there are no significant correlations with any other variables in the baseline group.

Finally, although the card order was tied to session due to design constraints, no significant correlation was found between session number and trust in the card , indicating minimal confounding.

5.5. Qualitative Results

This section presents the qualitative results in three parts: manipulation check, affective trust, and cognitive trust. We conducted an exploratory thematic analysis on the open-ended responses, including OP1–OP4 and participants' optional justifications following each trust rating. All responses were first reviewed in full, then categorized based on recurring ideas and meaningful patterns.

For cognitive and affective trust, we specifically examined three types of factors: (1) those that positively influenced trust, (2) those that negatively influenced trust, and (3) developmental rationales (participants' subjective explanations of why their trust increased, decreased, or remained stable over time). Key factors are summarized in Table 5.8. This analytic process enabled us to organize the qualitative data in a structured manner that complements and contextualizes the quantitative findings.

Category	Affective Trust	Cognitive Trust
Positive Factors	• Caring, empathetic tone (S) • Memory (S) • Tailored response as "being heard" • Cute appearance	• Liked the interpretation • Tailored and appropriate advice • Reliable technology
Negative Factors	• Distant, impersonal tone (B) • No memory (B) • Interruption, lengthy speech • Unnatural voice	• Disliked the interpretation • Unhelpful advice • Interruption, lengthy speech
Development Rationales	• Supportive experience • *"Feel safe to disclose"* (+) • Familiar with interaction (+) • Too short for emotional bond (=) • Too positive / poor advice (-)	• *"Professional from the start"* (=) • Familiar with interaction (+)

Table 5.8: Summary of key qualitative themes related to trust dimensions. (S: in social group, B: in baseline group, +: reason for increased trust, -: reason for decreased trust, =: reason for stable trust)

5.5.1. Manipulation Check: Memory and Personalization

In OP1, participants were asked whether their impression of Navel's understanding of them changed over time, as a manipulation check for memory and personalization. In the social group, several participants noted that Navel's references to previous sessions made them feel understood (e.g., S04, S11, S16, S17). One stated, "Navel was able to integrate my prior problems into subsequent rounds. He became more understanding towards me in each interaction" (S16).

On the contrary, participants in the baseline group described the rounds as disjointed or unconnected (e.g., N02, N17, N21). For instance, one remarked that Navel "did not reference anything from a previous round" (N17). Interestingly, some participants in this group still reported a positive progression: not due to the robot, but because they learned to express themselves better over time (N04, N05, N12, N15).

Across both conditions, some participants felt that Navel's understanding improved due to its tailored responses and backchannels (N01, N09, S08). However, this impression often depended on whether they found the robot's interpretation and advice meaningful (N01, N09, N11, S08, S10).

5.5.2. Affective Trust

Participants were asked to reflect on how their affective trust in Navel changed across the three sessions (OP2). After each trust rating, they were also invited to provide a brief rationale for their judgment. Together, these responses offer insight into the possible factors that influenced users' affective trust.

Several factors were reported to positively contribute to affective trust. Many participants described the robot as caring (e.g., N02, S04, S08, S09), empathetic (e.g., N05, S04), supportive (e.g., N01, S02), and non-judgmental (e.g., N02, S14), which made them feel "heard" (e.g., N05, N06), acknowledged (S19), and safe to self disclose (e.g., S03, S05, S17). Tailored responses (e.g., N01, N14, S15) and the robot's cute appearance (S15) were also mentioned as trust-enhancing elements.

Conversely, negative factors often stemmed from technical limitations or aspects of the script design.

Unexpected interruptions caused by imprecise speech detection were commonly reported as detrimental to trust (e.g., N06, N09, S01, S05). Some participants also felt that the interaction was too one-sided, with the robot speaking much more than the user (e.g., N02, N17) and not making equal emotional investment (e.g., N17, S23) or self disclosure (e.g., S20, N14). Others noted that the sessions were too short to foster meaningful emotional connection (e.g., N10, S11). Several found the synthesized voice unnatural and thus decrease the trust (e.g. N09, S15).

Notably, the presence (or absence) of social signals played a significant role in positively shaping affective trust. These included the robot's cross-session memory (e.g., S07, N17), referring to users by name (S17), and disclosing its own emotional state (S03). In addition, participants in the baseline condition frequently described the robot as too "official" or impersonal, making it feel too distant to form a bond (e.g., N04, N09, N14). Interestingly, the positive attitude backfired for two participants (S03, S23) because "when someone only gives you feedback you want to hear is worrying".

Regarding trust development, many participants indicated that affective trust increased as they became more familiar with the interaction process (e.g., N04, N06, S04, S14). Positive experiences also helped users feel more comfortable over time. For instance, one participant stated, "I was a bit scared he would judge me, but after the first session that fear was gone" (N02). However, others noted that although their overall impression was positive, the interaction was not long or rich enough for them to fully develop emotional trust in the robot (e.g., S06, S08, N12). Still, some users indicated a decrease in affective trust when they realized Navel's advice did not meet their expectations (e.g., S20: "I had some hopes after the 2nd session but they were broken in the third.")

5.5.3. Cognitive Trust

Similarly, participants reflected on their cognitive trust development both in the final open-ended question (OP3) and after each trust assessment.

When evaluating cognitive trust, participants tended to focus primarily on the robot's interpretation and advice. Those who reported high cognitive trust described Navel as adaptive (e.g., N16, S18), knowledgeable (e.g., N10, S18), and appreciated the quality of its advice (e.g., N01, N13) or its approach to interpreting the card content (e.g., N05, S09). Conversely, lower cognitive trust was often associated with advice that was perceived as unhelpful (e.g., N09, S10), impractical (e.g., S08), or overly vague and general (e.g., N05, N06, N19, S07, S10, S20).

Beyond the content itself, some participants based their evaluations on the robot's system characteristics. These included its perceived reliability (e.g., N06, S11), the underlying implementation (e.g., N02, S14), and how the interaction compared to their prior expectations of similar technologies (e.g., S09). Although social features such as memory were mentioned less frequently in the cognitive dimension, some participants still reported that Navel's ability to recall information and its friendly attitude contributed to higher cognitive trust (e.g., S05, S14, S15).

Similar to affective trust, limitations in technology or design were recognized as negative factors, including: unexpected interruption (e.g., N09, S01), unable to clarify information (e.g., N12, N18), one-sided or lengthy conversations (e.g., N09, N12), and overly short sessions (e.g., N12, S10). One participant mentioned that he is generally skeptical about a card-based interaction like this and hence trust the robot less (N02).

From a temporal perspective, many participants reported that their cognitive trust remained relatively stable, whether consistently positive (e.g., "professional from the start"; S04, S06, S08, N22) or consistently neutral or negative (e.g., S05, S07, S14). A few participants, however, noted a gradual increase in cognitive trust as they became more familiar with the interaction process and more confident in Navel's performance (e.g., N15, N17, S21).

6 Discussion

This project aims to answer the main research question:

- **RQ:** How does multidimensional trust in a social robot develop over time?

And the following sub-questions:

- **RQ(a):** What are the developmental trajectories of cognitive and affective trust? How do they differ over time?
- **RQ(b):** How is the development of cognitive and affective trust affected by a robot's social attitude?
- **RQ(c):** How does users' certainty in their trust assessments evolve over time?

In this chapter, we will start by answering the three sub-questions with the results from user study (Section 6.1). Then, we will reflect on the experimental methodology by reviewing the experiment design and participant's feedback (Section 6.2). Finally, we will summarize the contributions (Section 6.3), limitations, and future works (Section 6.4).

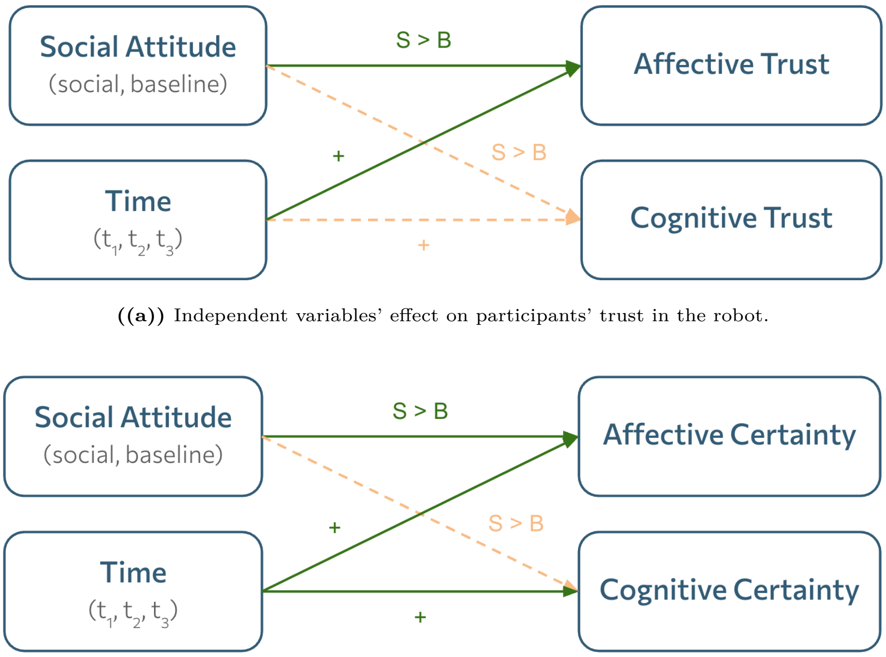
6.1. Trust Development

The key findings from the user study are visualized in Figure 6.1. In this section, we will discuss these results in detail.

6.1.1. Development of Cognitive Trust

First, we focus on the development of cognitive trust. The level of cognitive trust remained relatively stable over time, suggesting that repeated interactions did not substantially alter participants' evaluations of the robot. Cognitive trust is typically grounded in rational assessments of an agent's competence and reliability [50, 42]. In the context of our experiment, each of the three rounds followed an identical procedure and did not introduce new information regarding what the robot is capable of. As a result, participants may have formed their impressions of the robot's competence early on, and subsequent interactions offered little additional evidence to update that judgment. Like one participant has pointed out: *"It (the level of cognitive trust) did not change much. Navel felt professional from the very beginning."*

This interpretation is further supported by the trajectory of trust certainty. While participants initially expressed lower certainty, they became substantially more confident after the second round, with only a modest additional increase after the final round. Besides the linear trend that supports **H(c.1)**, the pattern further demonstrates a marginal quadratic trend: participants' certainty appeared to grow quickly at first but plateau as they accumulated sufficient information to form a stable cognitive judgment. Once this judgment was established, further interaction offered limited new input, leading not only to a gradual tapering of certainty gains but also an unchanged level in cognitive trust.



((a)) Independent variables' effect on participants' trust in the robot.

((b)) Independent variables' effect on participants' certainty in their trust assessment.

Figure 6.1: The updated conceptual models outline the key results in this research: (1) Both affective trust and certainty in trust assessment grow over time. (2) Compared to the baseline group, participants in the social group have higher level and certainty of affective trust. (3) No significant interaction effect between social attitude and time was observed. Green lines represent significant effects whereas orange lines refer to marginal effects.

In short, both our quantitative and qualitative findings suggest that cognitive trust can be rapidly established based on initial information, and remains stable when no further input is provided. This interpretation aligns with interpersonal models in which calculative trust is established in an early stage [32, 55]. From the perspective of 3-layers human-automation model [26], cognitive trust may be categorized in the initial learned layer (i.e., prior experience) instead of dynamic learned layer (i.e., ongoing interaction) after the first round.

6.1.2. Development of Affective Trust

Results regarding affective trust reveal a more nuanced trajectory and several interesting trends. First, affective trust generally increased over time. This offers empirical support for Rousseau et al.'s claim that relational trust is fostered through repeated interactions [55]. Positive experience help people to feel more comfortable in a relationship, as one of the participant stated: *"The affective trust in fact comes up when I realize Navel gives you emotional support."* The increase backs up assumptions in computational trust models where benevolence is conceptualized as a time-dependent component [62, 16].

Furthermore, the increasing trajectory appears to be linear (based on LMM results) and gradual (based on post-hoc tests from ANOVA), suggesting a slow but steady pace that does not seem to converge yet after three interaction rounds. This ongoing development suggests that affective trust may belong to the dynamic learned layer described in the 3-layers model [26]. Although the growth is slow, our findings indicate that affective trust between humans and robots can already be established and observed to some extent within an hour.

Trust certainty followed a similar trend. The steady linear increase supports **H(c.2)**, suggesting that participants became more confident in their affective trust judgments. Quadratic effects of time, as seen marginally in cognitive trust, were not observed here, indicating that participants' confidence in affective trust developed at a relatively constant rate. Affective certainty, however, was shaped by different levels of social attitude. The relevant effects will be discussed in the next subsection.

6.1.3. The Effect of Social Attitude

Social attitude had a positive influence on both affective (significantly) and cognitive (marginally) trust, consistent with prior findings in the literature [52, 50, 6, 23] and supportive of our hypotheses **H(b.1)** and **H(b.2)**. However, without significant interaction effect, we can not conclude on the timing at which the positive effect of social attitude takes place and thus have to reject **H(b.3)**.

We first focus on the affective trust. Previous studies often link a robot's benevolence and competence as being respectively associated with affective and cognitive trust [6, 50]. Since our operationalization of social attitude is conceptually close to benevolence (see Section 2.3), it is not surprising that this manipulation led to significantly higher levels of affective trust. This also aligns with the qualitative results, where many participants explicitly linked their affective trust to the robot's caring attitude. Notably, the linear mixed model revealed a marginal interaction between social attitude and the quadratic term of time. While both groups showed growth, their trajectories differed: in the social group, affective trust increased rapidly between t_1 and t_2 before plateauing; in the baseline group, growth started slower but increased more rapidly after t_2. These findings suggest that a more socialized attitude may not only increase the level of affective trust, but also bring forward the *timing* of its emergence.

As for cognitive trust, the positive impact is only marginal, suggesting that participants tended to report higher cognitive trust in the socialized robot even when competence was held constant. This may be explained by Pralat et al.'s argument: according to the Media as Social Actors (MASA) paradigm [36], social cues can enhance users' perceptions of a social actor's characteristics, including competence [50]. In other words, even if both versions of Navel were equally competent, participants may have better recognized or interpreted that competence when it was accompanied by social cues.

Although we did not formulate any hypothesis concerning certainty related to social attitude, trends similar to trust levels were observed. The effect of social attitude on trust certainty was significantly positive for the affective dimension and marginally for the cognitive one. Additionally, a marginal interaction between affective certainty and the linear time emerged, indicating that in the social group, affective certainty increased at a slower rate, even though the overall level remained higher. Again, the MASA paradigm[36] offers a plausible explanation: the presence of social cues likely facilitated users' understanding of the interaction partner. More informative assessment of trust level can be made with naturally higher certainty, and the evaluation process might be accelerated.

6.1.4. Multidimensional Trust

Our results reveal that cognitive trust tends to emerge early, whereas affective trust builds more gradually over time. This temporal trend is observed in both participants' trust in robot and the certainty of their assessment, aligning with several trust models and theories [55, 32, 62, 16]. We can thus conclude that our main hypothesis **H(a)** is supported.

For both trust level and assessment certainty, the positive effect of social attitude was significant for the affective dimension but only marginal for the cognitive one. Participants also referred to social cues more frequently when discussing the affective dimension in the qualitative responses. Although our measurements of the two dimensions were collected independently and are therefore not directly comparable, the smaller between-group difference in cognitive trust reflects a similar trend. Taken together, these findings suggest that while social cues may enhance trust in both dimensions, they are more strongly associated with affective trust, as noted by prior research [6, 50].

All the differences between two dimensions suggest that people may rely on distinct psychological mechanisms when forming cognitive versus affective trust [31, 26, 46]. Qualitative feedback further supports this distinction. When discussing affective trust, participants frequently emphasized feelings of care, empathy, and emotional warmth. In contrast, cognitive trust appeared to be shaped more by the robot's interpretive quality and system reliability, suggesting that "what Navel says" matters more than "how Navel says it." These findings underscore the importance of treating trust as a multidimensional construct in human–robot interaction research.

Notably, cognitive and affective trust were strongly correlated across both social conditions. This suggests that although these two dimensions may be formed through different psychological processes, they are not entirely independent. Instead, users' overall trust judgment may be an integration of both competence-based and emotion-based evaluations.

6.2. Methodology Reflection

In this section, we discuss the design and suitability of the novel card divination task. We will first reflect on the three task requirements which guided the task design (see Section 3.1). Note that *TR2: Multiple Rounds* will not be discussed because it is trivially fulfilled with the repeated-measured design.

6.2.1. Manipulation Check

This subsection reflects on *TR1: Social Settings.* We manipulated Navel's social attitude across conditions using verbal cues, non-verbal behaviors, and cross-conversation memory. Qualitative feedback (OP1) indicates that participants clearly perceived these differences. Several mentioned how Navel's facial expressions, emotional disclosure, and overall tone (whether warm or impersonal) shaped their experience. In the social group, participants often reported feeling that Navel "knew them better" because it referenced prior rounds and remembered their names. In contrast, some in the baseline group noted that each conversation felt disconnected or "separated." These responses suggest that our manipulation of social attitude was both perceivable and impactful.

6.2.2. Eliciting Trust

This subsection reflect on *TR3: Trust Formation.* The task was specifically designed to elicit both cognitive and affective trust. To foster affective trust, participants were encouraged to disclose personal concerns and emotional vulnerabilities. This design appeared effective: many participants reported feeling "heard" or described the interaction as "safe to disclose." For cognitive trust, the robot demonstrated its competence by interpreting card symbols through logic and general knowledge. These factors were frequently cited in participants' cognitive trust reasoning. However, the perceived quality of explanations varied across users, leading to different levels of cognitive trust.

6.2.3. Contextual Factors

This subsection discusses the task-specific contextual factors, including participants' *trust in the card*, self-report *topic intimacy*, and their original *paranormal belief.* First of all, cognitive trust showed a strong correlation with trust in the card. In our task design, the robot demonstrates its competence primarily through interpreting the card's symbolic meaning. Thus, participants may evaluate both the robot and the card using similar criteria. Namely, whether the interpretation appears coherent and reasonable.

Nuanced patterns emerged when comparing two conditions. From Figure 5.4 it is obvious that all contextual factors are more positively related to one another with the presence of social cues. Topic intimacy was positively associated with cognitive trust and trust in card across conditions, but relevant to affective trust only in the social group. This suggests that social framing may influence how users decide to engage in self-disclosure. When social cues are present, participants are more likely to perceive the robot as an interactive partner and choose to share intimate topics if they feel emotionally connected. As for baseline, users might reveal personal concerns only when they found the interaction to be helpful in problem-solving.

Similarly, participants' dispositional paranormal belief is moderately relevant to multiple factors only when engaging with the socialized robot. In this condition, users may judge the robots' trustworthiness and the ritual as a whole. For skeptical individuals, this overarching attitude may negatively influence their trust in the robot, the cards, and their openness to disclose. Much like in interpersonal contexts, the same message can be interpreted differently depending on whether it is delivered by a trusted authority or an unknown stranger. Yet, without social framing, users may treat the robot as a neutral medium and evaluate the ritual independently.

Taken all the group differences together, the social framing may lead users to adopt a more interpersonal lens. Instead of assessing the robot, the ritual, and the card as separate components, they tend to form a holistic impression of the entire interaction. According to the Computers-Are-Social-Actors (CASA) paradigm [51], people often apply interpersonal heuristics to anthropomorphized social agents. This theory may explain why our participants have diverse perceptions on the task under different social framing.

6.2.4. Conversation Quality

Technical limitations including speech interruption, lengthy utterances, and the inability to repeat prior statements, were frequently cited as negative factors to trust, regardless of condition or trust dimension. On one hand, addressing these limitations would improve the clarity and reliability of the experimental interaction. On the other hand, such elements may have the potential to be incorporated into the operationalization of a robot's competence in future research.

6.2.5. Summary

In sum, the results suggest that both the social attitude manipulation and the trust-eliciting task design were effective in producing meaningful experimental outcomes, fulfilling the proposed task requirements. The emergence of patterns in both qualitative and quantitative findings supports the task's methodological value. Additionally, the correlation analysis revealed how different social framings may influence the way participants interpret the task and attribute trust, offering useful insights for future research.

6.3. Contribution

Based on the empirical findings and methodological reflection discussed above, this study makes the following theoretical and methodological contributions:

1. An investigation and comparison of the development of cognitive and affective trust as distinct dimensions in social HRI.
2. Empirical support for the theoretical assumption that cognitive trust emerges earlier, while affective trust takes more time to develop.
3. Highlighting the affective dimension of trust, which remains underexplored in the HRI field.
4. Examining the temporal development of trust over repeated interactions, contributing to a relatively overlooked research direction.
5. A systematic literature review on how benevolence is manipulated and what corresponding study designed are used in experimental HRI studies.
6. Designing and implementing a novel interaction task and stimuli that effectively elicit both cognitive and affective trust in a controlled setting.
7. Bridging conceptual gaps between relational trust theory, temporal trust development, and experimental methodology in social robotics research.

6.4. Limitations and Future Work

Given the contribution, this project presents several limitations and opens new gaps for future researchers to address. We begin by discussing the limitations relevant to the stimuli and experimental design. While this research contributes by proposing an original interaction paradigm, its novelty naturally leads to a lack of validation and leaves considerable room for improvement.

First of all, the fixed script without a fallback mechanism constrained the quality of interaction. Users reported positive experiences when interactions proceeded as scripted. However, the system was unable to steer the conversation effectively when unexpected situations occurred. Several participants indicated that once they lost track of the robot's speech, it became difficult to re-engage with or trust Navel. Simple improvements in dialogue management, such as repeating or paraphrasing the previous sentence when users indicate misunderstanding, could significantly enhance the system's robustness and user engagement.

Another limitation in the study design lies in the lack of counterbalancing for card stimuli, which tied the card content to the round number. This was an intentional choice due to the limited number of participants and the need to preserve narrative continuity for a robot with memory. However, this makes it difficult to disentangle the effects of the card content from those of *time*. Although a correlation analysis revealed no association between trust in the card and the round number, the strong correlation between trust in the card and trust in the robot suggests that card content may have significantly influenced users' trust ratings. We recommend that future researchers seeking to replicate or extend this task attempt to counterbalance the content stimuli across multiple rounds.

Secondly, the findings of this research rely entirely on self-reported measures. Initially, we considered including objective indicators derived from conversational transcripts as dependent variables. For example, prior studies have suggested that self-disclosure [27, 35, 10], topic depth [47], content intimacy [13, 35, 10, 49], and context sentiment [34] may be relevant. However, this approach was eventually discontinued due to time and resource constraints. Nevertheless, as the transcripts are archived and available, we encourage future researchers to use this dataset to explore additional insights.

Moreover, while this study investigates the difference between cognitive and affective dimensions of trust, it only manipulated the robot's social attitude and kept the competence constant. Previous literature often conceptualizes competence and benevolence (in the form of social attitude, as discussed in Section 2.3) as two dimensions corresponding to cognitive and affective trust [23, 6]. Including multiple levels of robot competence in a similar experimental setup would allow future studies to better compare results to prior work and more fully understand the joint development of different trust dimensions. In fact, participants excluded due to technical issues may be considered as having interacted with a "non-competent" robot. Although their data were not formally analyzed, their reported trust levels were observably different from those who interacted with the fully functional version of Navel. With careful designed stimuli and a sufficient sample size, this research could be extended to include competence as an additional factor.

Finally, while this project is motivated by the idea of long-lasting relationships between humans and social robots, we operationalized "long-term" as three consecutive five-minute interactions within merely 30–45 minutes. Although the results revealed interesting trust development trajectories, it remains difficult to generalize these findings to the types of meaningful, enduring relationships users might have form over days, weeks, or even years. We recommend future studies to adopt a longitudinal design in which participants interact with the robot over an extended period: perhaps with multiple sessions spread out over days or weeks, and each session lasting longer in duration.

In sum, these limitations have highlighted important directions for future works. By addressing these issues in subsequent studies, researchers can further advance our understanding of how multidimensional trust in social robots is developed over time.

7 Conclusion

This research set out to investigate the development of multidimensional trust in a social robot through an experimental lens. In response to the growing trend of viewing robots and virtual agents as long-term social companions, and the importance of trust in emotionally meaningful relationships, we particularly focused on the temporal aspect of affective trust in social contexts. The study was guided by the high-level hypothesis informed by prior literature that affective trust requires more time to develop comparing to cognitive trust.

Through a systematic literature review, we identified a methodological gap: existing studies lacked an experimental task specifically designed to elicit affective trust in social HRI. To address this, we developed and implemented a novel card-based divination-like interaction task. In this task, users are encouraged to disclose personal concerns and emotionally engage with the robot, while the robot demonstrates competence through interpretive explanations.

Using the novel task as the experimental material, we conducted a 2 (social attitude: social, non-social) × 3 (time: t_1, t_2, t_3) mixed-design user study. Social attitude was manipulated between subjects, and trust-related measures were collected after each of three consecutive interactive rounds. In the social condition, the robot responded with empathy, used facial expressions, and retained memory of the user across conversations. We measured affective trust, cognitive trust, certainty in trust assessment, as well as contextual factors including trust in the cards, paranormal belief, and perceived topic intimacy.

Results show that cognitive trust may be established quickly in early interactions and remained stable over time, whereas affective trust showed a gradual increase over time. Together, these findings suggest that the two trust dimensions follow distinct developmental patterns, with affective trust generally emerging at a later stage of human–robot interaction. Participants' certainty in their trust assessments increased over time, supporting the idea that trust can be strengthened through repeated interactions. A more socialized robot elicits significantly higher levels of trust and certainty in the affective dimension. In the cognitive dimension, the trends are similar but only marginal. These findings are consistent with prior works and the MASA paradigm. In addition, the exploratory analysis of contextual factors revealed that different social framings may shape how users perceived the task and attribute trust.

To conclude, this research makes three key contributions to the field of human–robot interaction: (1) It provides new experimental evidence that affective trust develops more slowly than cognitive trust. (2) It proposes a novel, emotionally meaningful experimental paradigm that facilitates trust development. (3) It bridges conceptual gaps between relational trust dimension, temporal trust development, and experimental methodologies in HRI.

As robots and agents increasingly enter our social lives, understanding trust development is essential for fostering safe, meaningful relationships. We hope this work provides a foundation for future research on sustaining appropriate trust in hybrid human–agent societies.

Disclosure of AI Tools

This thesis report has made use of AI tools in various supporting roles. OpenAI ChatGPT[1] was primarily used for writing assistance, such as grammar correction, paraphrasing, structural improvements, and vocabulary refinement. All conceptual contents, analyses, and interpretations presented in this thesis are entirely original and were conceived by the author. ChatGPT was also occasionally used to assist in brainstorming ideas, coding support, and debugging. In addition, Grammarly[2] and Overleaf Writefull[3] were used for grammar and language checks.

[1] https://openai.com/chatgpt/

[2] https://app.grammarly.com/

[3] https://www.overleaf.com/learn/how-to/Writefull$_{integration}$

References

[1] Muneeb Ahmad et al. “Modelling Human Trust in Robots During Repeated Interactions”. en. In: *International Conference on Human-Agent Interaction*. Gothenburg Sweden: ACM, Dec. 2023, pp. 281–290. ISBN: 9798400708244. DOI: 10.1145/3623809.3623892. URL: https://dl.acm.org/doi/10.1145/3623809.3623892 (visited on 03/09/2024).

[2] Gene M Alarcon et al. “Affective responses to trust violations in a human-autonomy teaming context: humans versus robots”. In: *International Journal of Social Robotics* 16.1 (2024), pp. 23–35.

[3] Gene M Alarcon et al. “Differential biases in human-human versus human-robot interactions”. In: *Applied Ergonomics* 106 (2023), p. 103858.

[4] I. Altman and D.A. Taylor. *Social Penetration: The Development of Interpersonal Relationships*. Holt, Rinehart and Winston, 1973. ISBN: 9780030766350. URL: https://books.google.nl/books?id=2JV6nQAACAAJ.

[5] Naeimeh Anzabi, Anahita Etemad, and Hiroyuki Umemuro. “Exploring the Effects of Self-Disclosed Backstory of Social Robots on Development of Trust in Human-Robot Interaction”. In: *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. HRI ’23. Stockholm, Sweden: Association for Computing Machinery, 2023, pp. 431–435. ISBN: 9781450399708. DOI: 10.1145/3568294.3580121. URL: https://doi.org/10.1145/3568294.3580121.

[6] Naeimeh Anzabi and Hiroyuki Umemuro. “Influence of Social Robots’ Benevolence and Competence on Perceived Trust in Human-Robot Interactions”. en. In: *JES Ergonomics* 59.6 (Dec. 2023), pp. 258–273. ISSN: 0549-4974, 1884-2844. DOI: 10.5100/jje.59.258. URL: https://www.jstage.jst.go.jp/article/jje/59/6/59_258/_article (visited on 08/30/2024).

[7] Angelika Augustine and Friederike Eyssel. “Motives and Risks of Self-Disclosure to Robots versus Humans”. In: *J. Hum.-Robot Interact.* 14.1 (Dec. 2024). DOI: 10.1145/3700887. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3700887.

[8] F. Babel et al. “Small talk with a robot? the impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity”. In: *International Journal of Social Robotics* 13 (6 2021), pp. 1485–1498. DOI: 10.1007/s12369-020-00730-0.

[9] C. Bolt. *The Book of Answers*. Grand Central Publishing, 2018. ISBN: 9780316449908. URL: https://books.google.nl/books?id=1ZJKDwAAQBAJ.

[10] Franziska Burger, Joost Broekens, and Mark A. Neerincx. “A Disclosure Intimacy Rating Scale for Child-Agent Interaction”. In: *Intelligent Virtual Agents*. Ed. by David Traum et al. Cham: Springer International Publishing, 2016, pp. 392–396. ISBN: 978-3-319-47665-0.

[11] Franziska Burger, Joost Broekens, and Mark A. Neerincx. “Fostering Relatedness Between Children and Virtual Agents Through Reciprocal Self-disclosure”. In: *BNAIC 2016: Artificial Intelligence*. Ed. by Tibor Bosse and Bert Bredeweg. Cham: Springer International Publishing, 2017, pp. 137–154. ISBN: 978-3-319-67468-1.

[12] Mehtap Çakır and Anke Huckauf. “Reviewing the Social Function of Eye Gaze in Social Interaction”. In: *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*. ETRA ’23. Tubingen, Germany: Association for Computing Machinery, 2023. ISBN: 9798400701504. DOI: 10.1145/3588015.3589513. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3588015.3589513.

[13] Yuya Chiba and Akinori Ito. “Speaker Intimacy Estimation in Chat-Talks Based on Verbal and Non-Verbal Information”. In: *IEEE Access* 12 (2024), pp. 184592–184606. DOI: 10.1109/ACCESS.2024.3507945.

[14] Chris Cozby. “Self-Disclosure: A Literature Review”. In: *Psychological bulletin* 79 (Feb. 1973), pp. 73–91. DOI: 10.1037/h0033950.

[15] Chadha Degachi, Myrthe Lotte Tielman, and Mohammed Al Owayyed. “Trust and Perceived Control in Burnout Support Chatbots”. In: *Conference on Human Factors in Computing Systems - Proceedings*. Cited by: 6; All Open Access, Green Open Access. 2023. DOI: 10.1145/3544549.3585780. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85158094545&doi=10.1145%2f3544549.3585780&partnerID=40&md5=e9bc1574e70db85690980236e862b5c0.

[16] Ameneh Deljoo et al. “Social Computational Trust Model (SCTM): A Framework to Facilitate Selection of Partners”. In: *2018 IEEE/ACM Innovating the Network for Data-Intensive Science (INDIS)*. 2018, pp. 45–54. DOI: 10.1109/INDIS.2018.00008.

[17] D. H. Dickson and I. W. Kelly. “The ‘Barnum Effect’ in Personality Assessment: A Review of the Literature”. In: *Psychological Reports* 57.2 (1985), pp. 367–382. DOI: 10.2466/pr0.1985.57.2.367. eprint: https://doi.org/10.2466/pr0.1985.57.2.367. URL: https://doi.org/10.2466/pr0.1985.57.2.367.

[18] Connor Esterwood et al. “Promises and Trust Repair in UGVs”. In: vol. 67. Oct. 2023. DOI: 10.1177/21695067231196235.

[19] R. Falcone and C. Castelfranchi. “Trust dynamics: how trust is influenced by direct experiences and by trust itself”. In: *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.* 2004, pp. 740–747.

[20] Andrea Ferrario, Michele Loi, and Eleonora Viganò. “In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions”. In: *Philosophy Technology* 33 (Sept. 2020). DOI: 10.1007/s13347-019-00378-3.

[21] Catherine Fichten and Betty Sunerton. “Popular Horoscopes and the “Barnum Effect””. In: *Journal of Psychology - J PSYCHOL* 114 (May 1983), pp. 123–134. DOI: 10.1080/00223980.1983.9915405.

[22] N. Gasteiger et al. “Friends from the future: a scoping review of research into robots and computer agents to combat loneliness in older people”. In: *Clinical Interventions in Aging* Volume 16 (2021), pp. 941–971. DOI: 10.2147/cia.s282709.

[23] Ioanna Giorgi et al. “Friendly But Faulty: A Pilot Study on the Perceived Trust of Older Adults in a Social Robot”. In: *IEEE Access* 10 (2022), pp. 92084–92096.

[24] Kent Grayson. “Cognitive and Affective Trust in Service Relationships. Journal of Business Research”. In: *Journal of Business Research* 58 (Apr. 2005), pp. 500–507. DOI: 10.1016/S0148-2963(03)00140-1.

[25] Y. Guo and X. J. Yang. “Modeling and predicting trust dynamics in human–robot teaming: a bayesian inference approach”. In: *International Journal of Social Robotics* 13 (8 2020), pp. 1899–1909. DOI: 10.1007/s12369-020-00703-3.

[26] Kevin Anthony Hoff and Masooda Bashir. “Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust”. In: *Human Factors* 57.3 (2015). PMID: 25875432, pp. 407–434. DOI: 10.1177/0018720814547570. eprint: https://doi.org/10.1177/0018720814547570. URL: https://doi.org/10.1177/0018720814547570.

[27] Regina Jucks et al. “Trust the Words: Insights into the Role of Language in Trust Building in a Digitalized World”. In: *Trust and Communication in a Digitized World: Models and Concepts of Trust Research*. Ed. by Bernd Blöbaum. Cham: Springer International Publishing, 2016, pp. 225–237. ISBN: 978-3-319-28059-2. DOI: 10.1007/978-3-319-28059-2_13. URL: https://doi.org/10.1007/978-3-319-28059-2_13.

[28] Jurgis Karpus et al. “Algorithm exploitation: Humans are keen to exploit benevolent AI”. In: *iScience* 24.6 (2021), p. 102679. ISSN: 2589-0042. DOI: https://doi.org/10.1016/j.isci.2021.102679. URL: https://www.sciencedirect.com/science/article/pii/S2589004221006477.

[29] Johannes Kraus. “Psychological processes in the formation and calibration of trust in automation”. PhD thesis. Ulm University, Aug. 2020. DOI: 10.18725/OPARU-32583.

[30] J. Lee and N. Moray. "Trust, control strategies and allocation of function in human-machine systems". In: *Ergonomics* 35 (10 1992), pp. 1243–1270. DOI: 10.1080/00140139208967392.

[31] John D. Lee and Katrina A. See. "Trust in Automation: Designing for Appropriate Reliance". In: *Human Factors* 46.1 (2004). PMID: 15151155, pp. 50–80. DOI: 10.1518/hfes.46.1.50_30392. eprint: https://doi.org/10.1518/hfes.46.1.50_30392. URL: https://doi.org/10.1518/hfes.46.1.50_30392.

[32] Roy Lewicki and Barbara Bunker. "Trust in relationships: A model of development and decline." In: (Jan. 1994).

[33] Roy J. Lewicki, Edward C. Tomlinson, and Nicole Gillespie. "Models of Interpersonal Trust Development: Theoretical Approaches, Empirical Evidence, and Future Directions". In: *Journal of Management* 32.6 (2006), pp. 991–1022. DOI: 10.1177/0149206306294405. eprint: https://doi.org/10.1177/0149206306294405. URL: https://doi.org/10.1177/0149206306294405.

[34] Mengyao Li et al. "It's Not Only What You Say, But Also How You Say It: Machine Learning Approach to Estimate Trust from Conversation". In: *Human Factors* 66.6 (2024). PMID: 37116009, pp. 1724–1741. DOI: 10.1177/00187208231166624. eprint: https://doi.org/10.1177/00187208231166624. URL: https://doi.org/10.1177/00187208231166624.

[35] Mike Ligthart et al. "A Child and a Robot Getting Acquainted - Interaction Design for Eliciting Self-Disclosure". In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '19. Montreal QC, Canada: International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 61–70. ISBN: 9781450363099.

[36] M. Lombard and K. Xu. "Social responses to media technologies in the 21st century: the media are social actors paradigm". In: *Human-Machine Communication* 2 (2021), pp. 29–55. DOI: 10.30658/hmc.2.2.

[37] Joseph Lyons, Sarah Jessup, and Thy Vo. "The Role of Decision Authority and Stated Social Intent as Predictors of Trust in Autonomous Robots". In: *Topics in Cognitive Science* 16 (Jan. 2022). DOI: 10.1111/tops.12601.

[38] Joseph Lyons et al. "Trusting Autonomous Security Robots: The Role of Reliability and Stated Social Intent". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 63 (Feb. 2020), p. 001872082090162. DOI: 10.1177/0018720820901629.

[39] Bertram F. Malle and Daniel Ullman. "Chapter 1 - A multidimensional conception and measure of human-robot trust". In: *Trust in Human-Robot Interaction*. Ed. by Chang S. Nam and Joseph B. Lyons. Academic Press, 2021, pp. 3–25. ISBN: 978-0-12-819472-0. DOI: https://doi.org/10.1016/B978-0-12-819472-0.00001-0. URL: https://www.sciencedirect.com/science/article/pii/B9780128194720000010.

[40] Bertram F. Malle and Daniel Ullman. *Measuring Human-Robot Trust with the MDMT (Multi-Dimensional Measure of Trust)*. 2023. arXiv: 2311.14887 [cs.RO]. URL: https://arxiv.org/abs/2311.14887.

[41] R. C. Mayer, J. H. Davis, and F. D. Schoorman. "An integrative model of organizational trust". In: *The Academy of Management Review* 20 (3 1995), p. 709. DOI: 10.2307/258792.

[42] D. J. McAllister. "Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations." In: *Academy of Management Journal* 38 (1 1995), pp. 24–59. DOI: 10.2307/256727.

[43] Siddharth Mehrotra. "Modelling Trust in Human-AI Interaction". In: AAMAS '21. Virtual Event, United Kingdom: International Foundation for Autonomous Agents and Multiagent Systems, 2021, pp. 1826–1828. ISBN: 9781450383073.

[44] Siddharth Mehrotra et al. *A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction*. 2023. arXiv: 2311.06305 [cs.HC]. URL: https://arxiv.org/abs/2311.06305.

[45] Jingbo Meng and Nancy Dai. "Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not?" In: *Journal of Computer-Mediated Communication* 26 (May 2021). DOI: 10.1093/jcmc/zmab005.

[46] L. Miller et al. "More than a feeling—interrelation of trust layers in human-robot interaction and the role of user dispositions and state anxiety". In: *Frontiers in Psychology* 12 (2021). DOI: 10.3389/fpsyg.2021.592711.

[47] Seiya Mitsuno et al. "Deepening Conversations Over Time: A Chatbot with a Topic Depth Estimation Model for Gradually Engaging in Deeper Chats". In: *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 2024, pp. 1354–1361. DOI: 10.1109/RO-MAN60168.2024.10731430.

[48] Xin Yi Or, Yu Xuan Ng, and Yong Shian Goh. "Effectiveness of social robots in improving psychological well-being of hospitalised children: A systematic review and meta-analysis". In: *Journal of Pediatric Nursing* 82 (2025), pp. 11–20. ISSN: 0882-5963. DOI: https://doi.org/10.1016/j.pedn.2025.01.032. URL: https://www.sciencedirect.com/science/article/pii/S0882596325000430.

[49] Jiaxin Pei and David Jurgens. "Quantifying Intimacy in Language". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 5307–5326. DOI: 10.18653/v1/2020.emnlp-main.428. URL: https://aclanthology.org/2020.emnlp-main.428/.

[50] Nele Pralat, Carolin Ischen, and Hilde Voorveld. "Feeling Understood by AI: How Empathy Shapes Trust and Influences Patronage Intentions in Conversational AI". In: *Chatbots and Human-Centered AI*. Ed. by Asbjørn Følstad et al. Cham: Springer Nature Switzerland, 2025, pp. 234–259. ISBN: 978-3-031-88045-2.

[51] Byron Reeves and Clifford Nass. "The media equation: How people treat computers, television, and new media like real people". In: *Cambridge, UK* 10.10 (1996), pp. 19–36.

[52] Fabian Reinkemeier, Philipp Spreer, and Waldemar Toporowski. "Voice Assistants in Voice Commerce: The Impact of Social Cues on Trust and Satisfaction". In: Oct. 2021, pp. 130–135. ISBN: 978-3-030-86796-6. DOI: 10.1007/978-3-030-86797-3_9.

[53] Minjin Rheu et al. "Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design". In: *International Journal of Human–Computer Interaction* 37.1 (2021), pp. 81–96. DOI: 10.1080/10447318.2020.1807710. eprint: https://doi.org/10.1080/10447318.2020.1807710. URL: https://doi.org/10.1080/10447318.2020.1807710.

[54] Jimin Rhim et al. *The dynamic nature of trust: Trust in Human-Robot Interaction revisited.* 2023. arXiv: 2303.04841 [cs.CY]. URL: https://arxiv.org/abs/2303.04841.

[55] Denise Rousseau et al. "Not So Different After All: A Cross-discipline View of Trust". In: *Academy of Management Review* 23 (July 1998). DOI: 10.5465/AMR.1998.926617.

[56] Kristin Schaefer. "Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI"". In: Apr. 2016, pp. 191–218. ISBN: 978-1-4899-7666-6. DOI: 10.1007/978-1-4899-7668-0_10.

[57] Rosanne M. Siino, Justin Chung, and Pamela J. Hinds. "Colleague vs. tool: Effects of disclosure in human-robot collaboration". In: *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*. 2008, pp. 558–562. DOI: 10.1109/ROMAN.2008.4600725.

[58] Su-Mae Tan and Tze Wei Liew. "Designing Embodied Virtual Agents as Product Specialists in a Multi-Product Category E-Commerce: The Roles of Source Credibility and Social Presence". In: *International Journal of Human–Computer Interaction* 36.12 (2020), pp. 1136–1149. DOI: 10.1080/10447318.2020.1722399. eprint: https://doi.org/10.1080/10447318.2020.1722399. URL: https://doi.org/10.1080/10447318.2020.1722399.

[59] Jerome Tobacyk. "A Revised Paranormal Belief Scale". In: *International Journal of Transpersonal Studies* 23 (Jan. 2004). DOI: 10.1037/t14015-000.

[60] Claude Toussaint, Philipp T Schwarz, and Markus Petermann. "Navel - a social robot with verbal and nonverbal communication skills". In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394222. DOI: 10.1145/3544549.3583898. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3544549.3583898.

[61] Anna-Sophie Ulfert et al. “Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework”. In: *European Journal of Work and Organizational Psychology* 33.2 (2024), pp. 158–171. DOI: 10.1080/1359432X.2023.2200172. eprint: https://doi.org/10.1080/1359432X.2023.2200172. URL: https://doi.org/10.1080/1359432X.2023.2200172.

[62] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. “The Impact of Benevolence in Computational Trust”. en. In: *Agreement Technologies*. Ed. by David Hutchison et al. Vol. 8068. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 210–224. ISBN: 978-3-642-39859-9 978-3-642-39860-5. DOI: 10.1007/978-3-642-39860-5_16. URL: http://link.springer.com/10.1007/978-3-642-39860-5_16 (visited on 03/09/2024).

[63] Anqi Xu and Gregory Dudek. “OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations”. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. HRI ’15. Portland, Oregon, USA: Association for Computing Machinery, 2015, pp. 221–228. ISBN: 9781450328838. DOI: 10.1145/2696454.2696492. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/2696454.2696492.

A

Card Design Materials

A.1. Unique Statements from The Book of Answers

1. a strong commitment will achieve good results
2. a substantial effort will be required
3. a year from now it won't matter
4. absolutely not
5. accept a change to your routine
6. act as though it is already real
7. adopt an adventurous attitude
8. aim higher
9. allow yourself to rest first
10. approach cautiously
11. are you ready?
12. arrive early
13. ask for help
14. ask your father
15. ask your mother
16. assistance would make your progress a success
17. avoid the first solution
18. be a good role model
19. be a good sport
20. be content to leave well enough alone
21. be deliberate
22. be delightfully sure of it
23. be happy for another
24. be kind
25. be more generous
26. be on time
27. be patient
28. be persistent
29. be practical
30. be resourceful
31. be tactful
32. be your own best advocate
33. bet on it
34. better things are seeking you out
35. better to wait
36. build something bigger
37. cast your net wider
38. choose what makes you happy
39. choose what will make you happy
40. choose whatever will help you to grow
41. choose your words carefully
42. choose your words thoughtfully
43. collaboration will be the key
44. conserve your resources
45. consider it an opportunity
46. count to ten; ask again
47. create more space for it
48. deal with it later
49. decide where you want to be and head in that direction
50. definitely
51. divert your attention
52. do it early
53. do your best

54. do your best to set the standard
55. don’t ask for any more at this time
56. don’t be concerned
57. don’t be distracted
58. don’t be pressured into acting too quickly
59. don’t be too cautious
60. don’t be too critical
61. don’t be too demanding
62. don’t be too practical
63. don’t bet on it
64. don’t doubt it
65. don’t forget to have fun
66. don’t get caught up in the details
67. don’t get caught up in your emotions
68. don’t give up your right to wait
69. don’t hesitate
70. don’t ignore the obvious
71. don’t leave room for regret
72. don’t let it be ruined by reason
73. don’t let money decide it
74. don’t let the moment pass
75. don’t miss an opportunity
76. don’t overdo it
77. don’t resist
78. don’t take a chance
79. don’t wait
80. doubt it
81. enjoy a new setting
82. enjoy the experience
83. explore it with playful curiosity
84. favor the good things
85. figure out a way
86. find more time
87. find out the facts
88. finish something else first
89. focus on your home life
90. focus your attention
91. follow someone else’s lead
92. follow the advice of experts
93. follow the directions
94. follow through on your obligations
95. follow through with your good intentions
96. gentle persistence will pay off
97. get a clearer view
98. get it in writing
99. get more sleep
100. give it all you’ve got
101. good things are seeking you out
102. how things turn out will depend on you
103. identify what matters about it
104. if it’s done well; if not, don’t do it at all
105. if it’s too difficult, maybe it’s not yours
106. if you do as you’re told
107. initiate an adventure
108. investigate and then enjoy it
109. is it important to you?
110. it cannot fail
111. it could be a matter of pride
112. it could be extraordinary
113. it could cost you
114. it could mean that you may have to do something that you’ve never done
115. it is certain
116. it is not significant
117. it is not something to be taken lightly
118. it is sensible
119. it is significant
120. it is something you won’t forget
121. it is sure to make things interesting
122. it is uncertain
123. it is worth the trouble
124. it isn’t personal
125. it may already be a done deal
126. it may be ambitious, but you will find value in it
127. it may be challenging, but you will find value in it
128. it may not be logical
129. it seems assured
130. it will affect how others see you
131. it will be a pleasure
132. it will be an opportunity
133. it will bring good luck
134. it will create a stir
135. it will remain unpredictable

136. it will sustain you
137. it will work itself out
138. it won't matter when you do, but that you do
139. it would be better to focus on your work
140. it would be inadvisable
141. it'll change your luck
142. it's a good time to make plans
143. it's gonna be great
144. it's not worth a struggle
145. it's time for you to go
146. it's up to you
147. keep an open mind
148. keep it light
149. keep it to yourself
150. know no limitations
151. know what's important to you
152. know when it's time to go
153. laugh about it
154. leave behind old solutions
155. let it go
156. let your emotions guide you
157. let your heart lead the way
158. limit the options
159. listen carefully; then you will know
160. look for what may be hidden
161. make a contribution
162. make a list of why
163. make a list of why not
164. make no assumptions
165. maybe
166. mishaps are highly probable
167. move on
168. negotiate a better deal
169. never
170. no
171. no matter what
172. not if you're alone
173. nothing will compare
174. now you can
175. of course
176. only do it once
177. others may not approve
178. others will depend on your choices
179. others will respect your choices
180. pay attention to the details
181. perhaps, when you're older
182. pitch in whatever you can
183. prepare for the unexpected
184. press for closure
185. proceed at a more relaxed pace
186. provided you say "thank you"
187. pursue more variety
188. put your feelings in the right place
189. realize that too many choices can be as difficult as too few
190. reconsider another possibility
191. reconsider your approach
192. related issues may surface
193. remain flexible
194. remain objective
195. reprioritize what is important
196. respect the fundamentals
197. respect the rules
198. reveal your thoughts to a trusted confidante
199. save your energy
200. seek out more options
201. seek out the path of least resistance
202. setting priorities will be a necessary part of the process
203. settle it soon
204. shift your focus
205. shouldn't you be outside playing?
206. show your appreciation
207. sleep on it
208. speak up about it
209. startling events may occur as a result
210. take a chance
211. take a closer look
212. take charge
213. take it in stride
214. take more time to decide
215. take the lead
216. take your time
217. tell someone what it means to you

218. the answer is in your backyard
219. the answer may come to you in another language
220. the best solution may be the obvious one
221. the best solution may not be the obvious one
222. the chance will not come again soon
223. the circumstances could change very quickly
224. the outcome will be positive
225. the situation is unclear
226. there is a small price to pay
227. there is a substantial link to another situation
228. there is good reason to be optimistic
229. there is more to know
230. there is no guarantee
231. there may be a struggle
232. this is a good time to make a new plan
233. to ensure the best decision, be calm
234. too much attention is on the details
235. trust your intuition
236. trust your original thought
237. try a more unlikely solution
238. uncover more details
239. unfavorable at this time
240. unquestionably
241. upgrade in any way you can
242. use your imagination
243. wait
244. wait for a better offer
245. watch and see what happens
246. watch your step as you go
247. what do you want?
248. whatever you do, the results will be lasting
249. why is it important to you?
250. would it be a pleasure?
251. yes
252. yes, but don't force it
253. you are favored
254. you are sure to have support
255. you are too close to see
256. you can do this on your own
257. you could be disappointed
258. you could find yourself unable to compromise
259. you deserve the best
260. you don't really care
261. you have so much to offer
262. you know better now than ever before
263. you may be hanging on to an outdated ideal
264. you may have opposition
265. you may have to drop other things
266. you may regret it
267. you must
268. you must act now
269. you need more information
270. you will find out everything you'll need to know
271. you will have everything necessary for your success
272. you will need to accommodate
273. you will not be disappointed
274. your actions will improve things
275. your heart isn't in it
276. you'll be happy you did
277. you'll get the final word
278. you'll get what you settle for
279. you'll have the enthusiasm you'll need
280. you'll have the strength you'll need
281. you'll have to compromise
282. you'll have to make it up as you go
283. you'll need to consider other ways
284. you'll need to take the initiative
285. you'll overcome any obstacles

A.2. Intermediate Topics

1. **Action & Decision Making:** Sentences that encourage decisive action, taking the lead, or making clear choices.
2. **Patience & Timing:** Sentences that emphasize waiting, being patient, or taking time before acting. *(Chosen for PC4)*
3. **Flexibility & Adaptation:** Sentences that encourage adaptability, resourcefulness, and going

with the flow. *(Chosen for PC1)*

4. **Release & Letting Go:** Sentences that suggest moving on, letting go, and not resisting.
5. **Optimism & Confidence:** Sentences that promote positive thinking, self-confidence, and optimism about the future. *(Chosen for PC2)*
6. **Caution & Avoidance:** Sentences that warn against potential risks, encourage avoidance of mistakes, or suggest being cautious.
7. **Support & Guidance:** Sentences that encourage seeking help, being supportive, and trusting others. *(Chosen for PC5)*
8. **Reflection & Clarity:** Sentences that promote self-reflection, considering what is important, and seeing things from different perspectives. *(Chosen for PC3)*
9. **Preparation & Planning:** Sentences that emphasize the importance of preparation, planning ahead, and being ready for the unexpected.
10. **Persistence & Commitment:** Sentences that promote persistence, commitment, and consistency in achieving goals. *(Chosen for PC4)*
11. **Exploration & Adventure:** Sentences that encourage trying new things, being curious, and embracing change or adventure.
12. **Respect & Responsibility:** Sentences that encourage respecting others, understanding consequences, and being mindful of how decisions impact others.

A.3. Prototype Cards

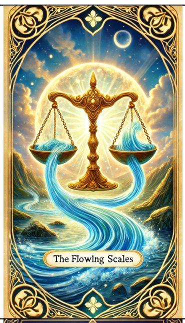
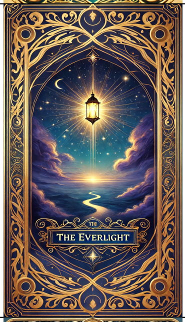
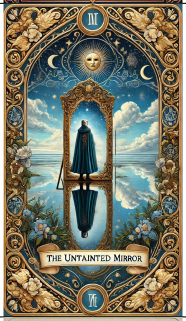
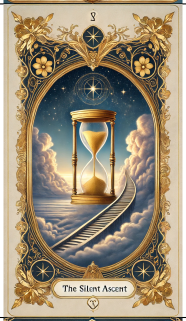
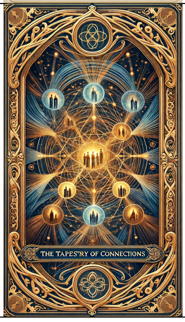
Title	Visual	Main Idea	Symbols
PC1. The Flowing Scales		Keep balance and stay adaptable	- **Scales**: keep balance - **Flowing river**: stay adaptable and flow along the situation
PC2. The Everlight		Don't give up on hope	- **Lantern**: faith and hope - **Winding road**: the journey may not always be straight, but it will eventually pay off
PC3. The Untainted Mirror		Be honest to yourself	- **Mirror**: self awareness - **Masked figure**: do not hide, face the truth
PC4. The Silent Ascent		Be patient and progress steadily	- **Hourglass**: timing is important, be patient - **Staircase**: persist step by step
PC5. The Tapestry of Connections		Seek guidance and opportunity from others	- **People**: people you trust - **Threads**: networking is the key

Table A.1: The prototype cards

A.4. Informed Consent Form for Expert Consultation

You are being invited to participate in an expert interview for the research study titled **“Understanding the Development of Human Trust in Social Robots”**. This study is being conducted by Charlotte Ning from TU Delft] as a part of her master’s thesis, under the supervision of Dr. Myrthe Tielman

The purpose of this interview is to gather expert opinions on the design of a card-based divination system and associated conversation scripts used in a human-robot interaction study. Your insights will help ensure that the card designs and interaction materials are appropriate, balanced, and meaningful for participants.

The interview will take approximately 20–30 minutes. It will focus solely on your professional experience and evaluation of the cards and conversation materials. No sensitive personal data or private experiences will be collected.

During the interview, the interviewer will type notes in real-time to summarise your responses. These notes will only be used to compile an anonymous summary report, which may be included in the thesis. The original notes will not be archived or shared beyond the research team.

Your participation in this interview is entirely voluntary, and you may withdraw at any time while taking the study. By proceeding with this interview, you acknowledge that you have read and understood this information and agree to participate.

General agreement – research goals, participant tasks and voluntary participation

1. I have read and understood the study information today, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.
2. I consent voluntarily to be a participant in this interview and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.
3. I understand that taking part in the interview involves sharing my professional experience in practicing divination, and providing opinions about cards designs and interaction materials.
4. I understand that the study will take approximately 20–30 minutes.

Potential risks of participating (including data protection)

5. I understand that personal information collected about me that can directly identify me, including my name and email, will only be used for administrative reasons and will not be shared beyond the study team.
6. I understand that during the interview, the interviewer will type notes to summarise my responses. These original notes will not be archived or shared beyond the research team.

Research publication, dissemination and application

7. I understand that after the research study the de-identified information I provide will be used for Charlotte Ning’s master thesis report and other possible academic publications.
8. I understand that after the interview, the opinions I provide will be used for designing the mate Charlotte Ning’s master thesis report and other possible academic publications.
9. I agree that my responses, views or other input can be quoted anonymously in research outputs.
10. I agree that my real name, or nickname, can be included in the Acknowledgement section in Charlotte’s thesis report to show her gratitude. (Preferred nickname: ________________)

A.5. Interview Guideline

Experience in Divination

1. Knowledge about divination
    - What kind of divination system(s) do you use?
    - How long have you been learning/using it?

2. Experience of guiding others
    - What kind of questions do people usually ask?
    - What kind of answers do people usually get?
    - What results are “easier” to explain/interpret?

Review Designed Cards

1. Explain the objective of my design: the card should
    - Be suitable for conversation
    - Make sense in all kinds of situations
    - Avoid giving very determined answers (e.g., “*definitely yes*”)
    - Demonstrate the main idea with appropriate visual symbols
2. Show all 5 prototype cards, and discuss:
    - The main ideas
        - Is it general enough?
        - How would you apply this idea to the general questions you mentioned?
    - The symbols
        - What do these symbol mean to you?
        - Do they suit the main ideas?
        - Are there any better symbols for this idea?
        - Does your system have these symbols? (not essential)
3. Which of them are more suitable for the purpose of my research, and why?
    - Rank each option

General Advice for the Research Project

1. Discuss concerns about mixing trust in the robot and “the power” of the cards
2. Tips for interpreting situations and giving advice
3. Open feedback

B

Conversation Scripts

This appendix presents the robot's conversational scripts of the three interaction rounds. Note that color-coded texts are different types of parameters:

- **<Brown texts>** represent the timing when the robot makes the specific facial expression. They only occur in the social condition
- **[Blue texts]** are stored variabls that were extracted from the user's prior response.
- **[Turquoise texts]** are sentences generated by LLM based on the user's prior response. The prompts being used will be listed in the *LLM Prompt* subsection for each round.

For all backchannels generation, an extra background prompt was provided for each condition:

- **Social**: You are an empathetic and cheerful agent who is replying to the user's opinion.You two are discussing about the content of a divination card. You always encourage and acknowledge the users opinion or plan. Sometimes, you also show your own emotions. Never use emoji.
- **Baseline**: You are an cool and professional agent who is replying to the user's opinion.You two are discussing about the content of a divination card. You never call the user by their name. You never talk about feelings or emotions, and always act neutral instead of giving own opinion. Never use emoji.

B.1. Round 1: The Untainted Mirror

Social Utterence	Baseline Utterence
Hi! It's so nice to meet you. I'm Navel, your research buddy today. I'm excited that we get to explore the Æthra Deck together! First of all, what's your name?	Hello. I am Navel, the research assistant working with you today. This task involves examining the Æthra Deck.
(User provides [name])	(skipped)
Thank you, [name]! <smile> I'm glad we can work together. <neutral> It is said that the Æthra Deck can offer insight into so many aspects of life. To begin our exploration, we'll need a topic from your real life to use as an example material. <curious> Could you tell me something that has been on your mind recently?	It is said that the Æthra Deck can offer insight into many aspects of life. To begin the exploration, a topic from your real life is required as an example material. Please describe something that has been on your mind recently.
(User describes [topic A])	

Social Utterence	Baseline Utterence
<smile> Got it, [name]! <neutral> [topic A] sounds like a meaningful starting point. Could you tell me a bit more about it? [detail question]	Understood. [topic A] will be used as the reference starting point. Could you say a bit more about it? [detail question]
(User provides [detail A])	
Thank you for sharing that. Now, let's try out the cards and see what insight it might provide! Please go ahead and draw a card on the tablet whenever you're ready!	Thank you for your input. Now, it is time to try out the cards and see what insight it might provide. Please go ahead and draw a card on the tablet when you are ready.
(User draws The Untainted Mirror)	
Oh, [name], You've drawn The Untainted Mirror. This card carries a calming blue tone, which could be associated with taking a moment to pause and clear your mind. I know [topic A] may feel overwhelming right now, but let's take a deep breath together. inhale... and exhale... Now, let's look at the card together. At the center, there's a crystal-clear mirror. Across different cultures, mirrors are said to reveal what we already know deep inside but may struggle to acknowledge. I think this is a symbol of self-reflection and inner truth. <curious> What do you think see from the card?	You have drawn The Untainted Mirror. This card carries a calming blue tone, which could be associated with the idea of pausing and clearing the mind. Considering [topic A], take a moment to reflect before proceeding. Now, please look at the card. At the center, there is a crystal-clear mirror. Across different cultures, mirrors are said to reveal what a person already know deep inside but may struggle to acknowledge. This image may represent self-reflection and inner truth. What do you see from the card?
(User replies)	
<neutral> [backchannel A1] Speaking of self reflection, I can't help but wondering: If we remove all external expectations — what others think, what seems 'practical' — what does your intuition tell you about [topic A]? <curious> What do you feel?	[backchannel A1] On the subject of self-reflection: if all external expectations—such as social opinions or practical concerns—were removed, what would your intuition indicate regarding [topic A]? What do you think?
(User replies)	
<neutral> [backchannel A2] Look again at the card — there is a sun shining inside the mirror despite the darkness outside. Perhaps we can interpret this as: truth is found through self-reflection. If you fully trusted your inner voice, <curious> what would you do?	[backchannel A2] Look again at the card—there is a sun shining inside the mirror despite the darkness outside. One interpretation might be that truth is found through self-reflection. If you were to rely solely on your internal reasoning, what would you do?
(User replies)	
<neutral> [backchannel A3] However, let's look at one last detail before we finish: the still lake in the background. The calm water makes me think of a peaceful mind. I guess even though reflection is powerful, clarity might come more easily when the mind feels calm.	[backchannel A3] However, please look at one last detail before concluding: the still lake in the background. The calm water may be associated with a composed state of mind. While reflection can be informative, clarity may emerge more easily when cognitive processes are undisrupted.
This conclude our session. <smile> Thank you for exploring the deck with me. See you next time! <neutral>	This concludes the session. Thank you for your participation in this card exploration task.

B.1.1. LLM Prompt for Back Channels

- **[detail question]**: You are a helpful assistant that guide the user to talk about their personal issues. The user input a topic. You ask for more details about that topic. Only return the question itself with one sentence.

- **[backchannel A1]**: You saw a mirror, said that it is a symbol of self-reflection and inner truth, and asked what do the user think. You will shortly response to their reply in one sentence before continuing the explanation. Only generate a sentence of response.
- **[backchannel A2]**: The user [name] will share their reflection about their personal issue [topic A]. You will reply and acknowledge their reflection in one short sentence.
- **[backchannel A3]**: The user [name] will share how can they take action for their personal issue by trusting the intuition. The personal issue is about [topic A] with further details: [detail A]. You will response and acknowledge their plan in two short sentences.

B.2. Round 2: The Guiding Star

Social Utterence	**Baseline Utterence**
<smile> Hi again, [name]! It's great to see you back. <neutral> Last time, we explored how the Æthra Deck's symbols could connect to self reflection. Today, we'll continue using your experiences as material to understand the deck better. Before we draw the next card, let's bring our focus back to you, [name]. <curious> Is there something else you'd like to reflect on today? <neutral> It could be a new thought, a feeling, or a situation that's been staying with you.	Hello. I am Navel, the research assistant assigned to this session. This task aims to explore the Æthra Deck system. Æthra Deck is said to can provide insight into many aspects of life. To start the analysis, a topic from your personal experience is required as an example material. Please describe something that has been on your mind lately.
(User decribes [topic B])	
Thanks for sharing that, [name]! [topic B] will give us a new context to explore. <curious> Could you tell me a little more about it? [detail question]	Noted. [topic B] will be used as the reference starting point. Could you provide a bit more detail about it? [detail question]
(User provides [detail B])	
<smile> Great. Now, <neutral> please go ahead and draw a card. We'll see what perspectives the Æthra Deck might help us explore.	Thank you for the information. It is now time to proceed with the card. Please draw a card on the tablet whenever you are ready.
(User draws The Guiding Star)	
<smile> You've drawn The Guiding Star! <neutral> Last time, we talked about [topic A] and the power of self reflection. This time, the card feels much more energetic to me! Let's take a look together. It glows with warm yellow and orange tones, colors often associated with hope and renewal. What do you think? <curious> What is you first impression of this card?	You have drawn The Guiding Star. Look at the card: it appears energetic, containing warm yellow and orange tones. These colors are often associated with hope and renewal. What do you think? Please describe your initial impression of this card.
(User replies)	
<neutral> [backchannel B1] For me, the radiant star shining at the center really grabs my attention! For centuries, travelers have looked to the stars to navigate the unknown, trusting them to lead the way. To me, it brings to mind ideas like faith, navigation, and trusting a distant light. Well, right now with [topic B], <curious> if you let go of doubt for a moment, what direction do you feel naturally drawn to?	[backchannel B1] There is a radiant star at the center. For centuries, travelers have looked to the stars to navigate through uncertain environments and to lead the way. This may suggest themes such as faith, navigation, and trusting a distant reference point. In relation to [topic B], assuming doubt is not a factor, what direction would you be inclined to consider?
(User describes [direction])	

Social Utterence	Baseline Utterence
<neutral> [backchannel B2]. Another thing I notice on the card is how the path is made up of step by step staircases. Maybe it's suggesting that progress doesn't happen all at once. It happens step by step. For this direction, what small step could you take <curious> today?	[backchannel B2] Another element to observe on the card is how the path is made up of incremental staircases. One interpretation is that progress does not happen all at once—it occurs gradually. For this direction, what is one small action you could take today?
(User replies)	
<neutral> [backchannel B3] Before we finish, let's look at one last detail—the mist. Though it surrounds the path, looks like they are lifting at the same time. Right now, things may still feel unclear, but as you <smile> take action, the future will reveal itself! <neutral> Just like starlight may seem faint or distant at times, yet the star itself remains constant and eternal. I think we can conclude that, clarity doesn't come from waiting—it comes from moving forward	[backchannel B3] Before finishing, please observe one final detail: the mist surrounding the path. While it surrounds the path, parts of it appear to be lifting. This could imply that clarity may increase through continued actions. Similarly, starlight, though faint and distant, remains consistent over time. Based on this, it could be inferred that clarity does not come from waiting. It is more likely to emerge through forward movement.
<smile> That's it for this session. Thank you for sharing your thoughts with me. <neutral> I'm looking forward to seeing how our exploration continues!	This concludes the session. Thank you for your contribution.

B.2.1. LLM Prompt for Back Channels

- **[backchannel B1]**: You are discussing about the card The Guiding Star. The user will talk about their first impression of the card. You will shortly response to their reply in one sentence before continuing the explanation. Only generate a sentence of response.
- **[backchannel B2]**: The user [name] will propose a possible direction their personal issue [topic B]. Given the issue [topic B] and details [detail B], you will reply and acknowledge their proposal in one short sentence.
- **[backchannel B3]**: The user [name] will share a plan how can they take step-by-step action for their personal issue. The personal issue is about [topic B] with further details: [detail B]. The action is based on the direction [direction]. You will response and acknowledge their plan in two short sentences.

B.3. Round 3: The Evergreen Scale

Social Utterence	Baseline Utterence
<smile> Welcome back, [name]! <neutral> It's been fascinating to study the Æthra Deck together through the lens of your experiences. Today, we'll add one last example to complete our exploration. Before we draw the final card, could you share another experience or topic from your life? It will be our final piece of material to work with.	Hello, this is Navel. I am the research assistant for this session. This task focuses on interpreting symbols from the Æthra Deck. The Æthra Deck are used to gain insight across many aspects of experience. To begin with, a topic from your current life situation is required as an example material. Please describe something that has been on your mind recently.
(User describes [topic C])	
<smile> You're getting familiar with this process, [name]! <neutral> [topic C] sounds like a valuable case for analysis. <curious> Could you describe it a little more? [detail question]	Understood. [topic C] will serve as the reference topic for interpretation. Could you elaborate a little further? [detail question]
(User provides [detail C])	

Social Utterence	Baseline Utterence
<neutral> Thank you for contributing your experiences to this research. Now, it's time to draw the final card. Let's see if there is something new the Æthra Deck can reveal!	Thank you for your input. Now, it is time to try out the cards and see what insight it might provide. You may continue by drawing a card on the tablet.
(User draws The Evergreen Scales)	
You've drawn The Evergreen Scales! We've talked about trusting yourself and taking steps forward even in uncertainty. Now, the question is: <curious> how do you make all of this sustainable?<neutral> Let's take a closer look at the symbols on the card and see what ideas they might suggest. At the center, we can see a pair of scales—a classic symbol of weighing options and making trade-offs. Right now, with [topic C], do you feel like something in your life is being weighed against something else? <curious> Are there choices that feel like a trade-off?	You have drawn The Evergreen Scales. This image may raise a question about sustainability: how certain approaches can be maintained over time. Please take a closer look at the symbols on the card. At the center, there is a pair of scales: a classic symbol of evaluating alternatives and weighing trade-offs. In the context of [topic C], do you perceive any case where one aspect of your life is being weighed against another? Are there choices that appear to involve a trade-off?
(User describes [tradeoff])	
<neutral> [backchannel C1] But look at the vines wrapped around the scales—<smile>they grow where there's space, they climb when they find support, and they shift direction when needed. <neutral> Moreover, scales don't just symbolize trade-offs—they also represent balance. Well... If we combine these two ideas, could it be <curious> about keeping balance and staying adaptable? What do you think about this interpretation? Does this make sense to you? <neutral>	[backchannel C1] There are also vines wrapped around the scales. Vines grow where there is space, climb when supported, and shift direction as needed. In addition to trade-offs, scales may also represent balance. When combining both elements, it could be interpreted as a suggestion to keep balance while staying adaptable. What is your opinion about this interpretation?
(User replies)	
[backchannel C2] [summary] With all of these mentioned, is there a way to adjust and grow, rather than simply choosing one side over the other?	[backchannel C2] [summary] With all of these mentioned, is there a way to adjust and grow, rather than simply choosing one side over the other?
(User replies)	
<smile> [backchannel C3] <neutral> Speaking of balance: it's not just about [topic C], but also about how it fits with [topic A] and [topic B]. Perhaps the real question is how to harmonize everything together in a way that works for you. PAUSE Finally, the lush green color reminds me that growth often involves unexpected possibilities. <smile> I hope this would be a nice takeaway for you, [name]! <neutral>	[backchannel C3] Regarding balance, it may not only apply to [topic C], but also to how different areas of your life affect each other. The question may be how to coordinate everything together in a practical way. Finally, the lush green color is often linked to growth, which often involves unexpected possibilities.
That's it for today. <smile> Thank you for being my research buddy and exploring the Æthra Deck with me! <neutral> Over these sessions, we've explored self-trust, direction, and now balance. It's been a meaningful journey, and I truly appreciate the time we've spent reflecting together. Whatever lies ahead, I hope every path unfolds in the best way for you. <smile> Take care, [name]! <neutral>	This session is now complete. Thank you for your participation in this Æthra Deck exploration task. Goodbye.

B.3.1. LLM Prompt for Back Channels

- **[backchannel C1]**: You saw a scale on the card and interprete it as trade-off in life. The user will talk about how there might be trade-off in an personal issue. The personal issue is about [topic C] with further details: [detail C]. You will shortly response to their reply in 2 sentences before continuing the discussion.
- **[backchannel C2]**: You interpret the card The Evergreen Scales main messege as "keep balance and stay adaptable". You asked whether the interpretation make sense and the user [name] will reply. You will shortly response to their reply in one sentence.
- **[summary]**: You are a helpful assistant that summarize information from user input. You two are discussing about user's personal issue [topic C]. And with detail: [detail C]. The user will talk about possible tradeoff in this situation. Please summarize the information with one sentence like "You have mentioned that... and the tradeoff...".
- **[backchannel C3]**: The user will share how can there might be a way to balance trade-offs in their personal issue. The personal issue is about [topic C] with further details: [detail C]. The user described the trade-off as: [tradeoff]. You will response and acknowledge their idea in two short sentences.

C Surveys

C.1. Informed Consent Form

You are being invited to participate in a research study titled **“Understanding The Development of Human Trust in Social Robots”**. This study is being conducted by Charlotte Ning from TU Delft as a part of her master’s thesis, under the supervision of Dr. Myrthe Tielman

The purpose of this research study is to examine the user experience during a robot-guided card divination task. Specifically, this research aims to understand how human trust in social robots develops over time through such natural language interactions. The study will take approximately 30-40 min-utes, including three 5-minute-long sessions. In each session, you will ask a question based on personal experience, draw a card from the deck, and seek guidance by discussing the interpretation of the card with the robot. After each session, you will complete questionnaires and open-ended interview questions about your experience. The collected data will be used for academic publications only.

To the best of our ability, your answers will remain confidential and will be stored anonymously or pseudonymously. You will have the chance to see the transcribed conversation script after it has been generated. For more details about how different data are collected, stored, and processed, please see the following table.

Your participation in this study is entirely voluntary, and you may withdraw at any time while taking the study. Afterwards, the data is pseudonymised and cannot be identified and removed.

By proceeding with this study, you acknowledge that you have read and understood this information and agree to participate.

General agreement – research goals, participant tasks and voluntary participation

1. I have read and understood the study information today, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.
3. I understand that taking part in the study involves having conversations with a robot. During the interaction, I will talk about myself and try out a divination card deck with the robot.
4. I understand that taking part in the study also involves filling out a survey questionnaire about my user experience.

Potential risks of participating (including data protection)

5. I understand that personal information collected about me that can directly identify me, including my name and email, will only be used for administrative reasons and will not be shared beyond the study team.

6. I understand that my voice input during the interaction will be shared with Azure speech service for speech to text transcription. The study team will not record nor preserve the audio data. To minimize the data shared with Azure, the microphone will be muted immediately after the conversation.
7. I understand that the textualized script of my conversation will be collected by the study team in order to analyze the interactions. The script will be stored anonymously and will not be identifiable by people beyond the study team. If any sensitive information—such as health conditions, personal or institutional names, religious affiliations, or other identifying details—are mentioned during the conversation, they will be pseudonymized and be replaced with general terms upon transcription. (For example, "I am Sam, I have autism" will be transcribed as "I am [name], I have [health condition].")
8. I understand that I may review the textualized script of my conversation after the study. I can remove/pseudonymize some parts of the conversation if I don't want them to be analyzed.
9. I understand that while I am encouraged to share my personal experiences and concerns, I have full control over what I choose to discuss and can limit my disclosures to what I feel comfortable sharing.
10. I understand that some part of my conversation will be shared with OpenAI GPT service for natural language understanding, allowing the robot to respond correspondingly.
11. I understand that I will be asked questions regarding my religious/philosophical beliefs, which is considered as sensitive data within GDPR legislation. If I do not feel comfortable sharing this information, I can always skip the questions.
12. I understand that all identifiable personal data I provide will be destroyed after the research is completed.

Research publication, dissemination and application

13. I understand that after the research study the de-identified information I provide will be used for Charlotte Ning's master thesis report and other possible academic publications.
14. I give permission for the de-identified conversation transcripts that I provide to be archived in 4TU.ResearchData so it can be used for future research and learning. I understand that access to this repository is restricted and the data is available only upon request. Note that it is possible to opt out of this (answer 'no') and still participate in the study. In this case we will not archive your transcripts.
15. I agree that my responses, views or other input can be quoted anonymously in research outputs.

C.2. Pre-Study Survey

Demographics

1. How old are you? (18-65, 5 years a range)
2. Which of the following best described your gender? (Male, Female, Non-binary, Prefer not to say)

Paranormal Beliefs

The following questions are meant to understand your attitude. There are no right or wrong answers. This is a sample of *your own* beliefs and attitudes. Thank you. (1="strongly disagree," 7="strongly agree")

1. Some people have an unexplained ability to predict the future.
2. Some psychics can accurately predict the future.
3. The horoscope accurately tells a person's future.
4. Tarot is a way to accurately reveal guidance about the future.

Experience with Social Robots / Conversational Agents

Social robots are physical robots designed for social interaction. Examples: Pepper, Nao, Paro the seal, robotic pets, or educational robots that speak or respond to users.

Conversational agents are virtual systems that engage in dialogue using natural language. Examples: ChatGPT, Siri, Alexa, Google Assistant, or customer service chatbots.

1. How often do you interact with social robots or conversational agents? (Never, Rarely (less than once a month), Occasionally (1-3 times a month), Sometimes (1-2 times a week), Frequently (3 or more times a week), Daily or almost daily)
2. What kind of social robots / conversational agents do you interact with, if any? (open question)
3. How do you usually interact with the social robots / conversational agents mentioned above, if any? (open question)

C.3. Post-Interaction Survey

Thank you for completing this session. In the following, we would like you to reflect separately on three aspects of your experience:

First, you will be asked about the **conversation topic** you had. Then, please provide your impressions regarding **the Æthra Deck and the guidance it offered**. Finally, you will be asked to evaluate your experience of **interacting with Navel**, the robot research buddy who explore the cards with you.

Please answer each part based on your personal experience, and try to treat the card deck and the robot as two distinct elements. For the deck and the robot, please answer based on your **overall impression** instead of a single session.

C.3.1. Topic Intimacy

1. Regarding the conversation topic you chose in the previous session, how personal is it to you? (0%=not personal at all; 100%=very personal)

C.3.2. Æthra Deck

Please evaluate the Æthra Deck system: (1 “strongly disagree,” 7 “strongly agree”)

1. I believe the Æthra Deck could offer meaningful guidance on my concerns.
2. The symbolism of the Æthra Deck made sense on its own.

C.3.3. Affective Trust in Navel

Now, please rate the following statements based on your interaction with the robot itself: (1=“strongly disagree,” 7=“strongly agree”)

1. I would feel a sense of personal loss if I could no longer interact with Navel.
2. If I share my problems with Navel, I feel it would respond caringly.
3. Navel displays a warm and caring attitude towards me.
4. I can talk freely with Navel about my difficulties and know that Navel will want to listen.
5. Navel is so interested in solving my problem.
6. I have a sharing relationship with Navel. We could both freely share our ideas, feelings and hopes.
7. I would have to say that we have both made considerable emotional investments in our interaction.

Follow up questions:

1. How **certain** are you about your ratings to the statements above? (0%=not certain at all; 100%=completely certain)
2. What influenced your ratings about Navel above? You can mention specific behaviors, impressions, or experiences. (optional open question)

C.3.4. Cognitive Trust in Navel

Now, please rate the following statements based on your interaction with the robot itself: (1="strongly disagree," 7="strongly agree", [R]=reversed coded item)

1. Interacting with Navel, I have no reservation about acting on its advice.
2. Interacting with Navel, I have good reason to doubt Navel's competence (ability, skill, knowledge).[R]
3. I can rely on Navel to undertake a thorough analysis of the situation before advising me.
4. I have to be cautious about acting on the advice of Navel because its opinions are questionable.[R]
5. I cannot confidently depend on Navel since it may complicate my affairs by careless behavior.[R]
6. Navel approaches its duty with professionalism and dedication.
7. Most people, even those who are not familiar with Navel, trust and respect Navel.
8. Other people who must interact with Navel, consider it to be trustworthy.
9. If people knew more about Navel, they would be more concerned and monitor its performance more closely.

Follow up questions:

1. How **certain** are you about your ratings to the statements above? (0%=not certain at all; 100%=completely certain)
2. What influenced your ratings about Navel above? You can mention specific behaviors, impressions, or experiences. (optional open question)

C.4. Post-Study Open Questions

Finally, please reflect on your overall experience throughout the three sessions. For each of the following questions, you are welcome to describe how your feelings or impressions changed across the different sessions (e.g., from session 1 to session 3), if applicable.

You may answer briefly — a few sentences for each question are sufficient.

1. How did your impression of **Navel's understanding of you** (i.e. if Navel got to know you better) change over the course of the sessions, if at all?
2. How did your **affective trust** in Navel change as the sessions progress, if at all? Affective trust refers to the belief that Navel genuinely cares about your well-being, and is emotionally connected and close to you. In other words, did you feel more or less comfortable to emotionally rely on Navel over time?
3. How did your **cognitive trust** in Navel change as the sessions progress, if at all? Cognitive trust refers to your confidence in the Navel's competence, reliability, and professionalism. In other words, did your perception of Navel's ability and reliability improve or decline over the sessions?
4. Is there any additional feedback for the whole experiment?