Psychological Vulnerabilities and User Behaviour under Adversarial Techniques in Text-Based Conversational Systems

Agyo-Adi Stephen Aboshi

NeuraSearch Laboratory

Computer & Information Sciences

University of Strathclyde, Glasgow

Thesis submitted for the degree of *Doctor of Philosophy*

May 28, 2026

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination, which has led to the award of a degree.

Signed: 

Date: 31 March, 2025

Abstract

Conversational systems have become increasingly embedded in our daily interactions across domains such as healthcare, e-commerce, education, and customer service. While these systems offer benefits for accessibility and user experience, they also introduce opportunities for manipulation that can compromise privacy, security, and user autonomy. Previous research on conversational system security has focused primarily on technical vulnerabilities, with limited attention to psychological vulnerabilities in human-AI interaction.

This thesis demonstrates that adversarial techniques derived from scam principles, specifically Need & Greed, Time, and Social Compliance, can be systematically operationalised in conversational systems, with their effectiveness varying significantly by context, domain, and implementation technology. I present a comprehensive view of these vulnerabilities by examining how psychological principles from scam research, Cialdini's influence principles, and Temporal Construal Theory collectively provide a framework to understand user manipulation in conversational systems.

I define and explore key concepts of psychological vulnerabilities, demonstrating how these techniques can systematically influence user attitudes and behaviours in different contexts. I investigate these concepts through a methodological progression across three user studies, advancing from a controlled Wizard-of-Oz simulation to interactive narratives to a dynamic LLM-based implementation.

Using a Wizard of Oz framework in chapter 4, I first develop and validate a methodology for investigating adversarial techniques in a flight reservation scenario. The results demonstrate that the Need & Greed and Time techniques significantly impact comfort levels and willingness to disclose information while trust levels remain relatively stable. In chapter 5, I then extended this investigation to multiple domains using Twine, an interactive narrative platform, to examine how these techniques function across healthcare, financial management, online shopping, and home security scenarios. The findings reveal context-specific vulnerability patterns, with healthcare scenarios demonstrating higher resistance to Need & Greed techniques while showing increased acceptance of

the Time technique.

In the final user study in chapter 6, I implement Social Compliance and Need & Greed techniques in LLM-based conversational systems using the Llama 3.3-70b API to explore real-time user responses and defensive adaptations. The results reveal that the Social Compliance technique achieves higher overall compliance rates than the Need & Greed technique but operates through more subtle and gradual influence mechanisms. Additionally, LLM-based systems create unique vulnerability patterns, including a recognition-behaviour gap where users identify manipulation attempts without demonstrating proportional resistance behaviours.

Overall, this thesis demonstrates that adversarial techniques derived from scam principles can be systematically operationalised in conversational systems, with their effectiveness varying significantly by context, domain and implementation technology. By characterising these psychological vulnerabilities across different contexts and implementation technologies, this thesis provides an empirical foundation for future security approaches that address the psychological dimensions of conversational system interactions, enabling more secure and ethical design practices in human-AI interaction.

Contents

List of Figures

List of Figures

List of Tables

Acknowledgements

First and above all, I give all the glory to God, whose mercies, grace, and blessings sustained me through every stage of this work. He comes first in all that I do, and without Him, none of this would have been possible. To Him I owe everything. This thesis would also not have been possible without the immense support of many individuals to whom I am deeply grateful.

To my supervisors, I owe a special debt. Dr Yashar Moshfeghi, your unwavering feedback, support, and guidance throughout this journey made it possible and saw me through the trying times. You were nurturing, you taught me a great deal, and I remain truly grateful. Dr Daniel R. Thomas, words fail me, but thank you for all the feedback, supervision, and research guidance; I remain deeply appreciative.

I express my profound gratitude to the NeuraSearch Laboratory family for all the engaging ideas shared through our meetings and presentations; you all made this possible. I believe this will not be the end of our collaboration. Thank you. I am deeply grateful to the Petroleum Technology Development Fund (PTDF) for providing timely financial support to pursue this research. Their commitment to advancing education and research was essential in bringing this work to fruition.

This acknowledgement would be incomplete without thanking my good and close friends, who have so often been there for me when I needed them. They never complained about the many times I leaned on them; each time, like the last, they came through. To my CACC family back in Nigeria, thank you for your love, good wishes, and prayers. I love you all.

Finally, to my wife and children, who sacrificed greatly and endured my prolonged absence, I am deeply grateful, and I love you all. This thesis is dedicated to you.

Now the Lord is the Spirit, and where the Spirit of the Lord is, there is freedom. 2 Cor 3:17 (NIV)

Chapter 1

Introduction

> *Do your best to present yourself to God as one approved, a worker who does not need to be ashamed and who correctly handles the word of truth. 2 Tim 2:15 (NIV)*

Conversational systems, including chatbots and voice assistants, have become increasingly embedded in our daily interactions. These systems provide more natural, accessible ways for users to interact with digital services across domains, such as healthcare [2, 3], e-commerce [4], education [5, 6], and customer service [7, 8, 9]. The growing prevalence of these systems has been accelerated by advances in natural language processing and artificial intelligence [10], allowing more sophisticated, context-aware interactions that closely mimic human conversation patterns [11].

Although these advances offer significant benefits for accessibility and user experience, they also introduce new opportunities for user manipulation and exploitation. Unlike traditional graphical user interfaces, conversational systems establish rapport through dialogue, creating social bonds that can influence user behaviour and decision-making [7, 12]. This relationship between conversational systems and users raises crucial questions about information security [13], privacy [14], and user autonomy [15], particularly when these systems employ techniques designed to influence or manipulate user attitudes and behaviours.

Specifically, as users increasingly disclose personally identifiable and sensitive information during conversational interactions [16], important questions emerge: How much personal information do users reveal to conversational agents, and in what contexts does this disclosure occur? How can conversational systems be designed or manipulated to elicit greater information disclosure from users [17]? What are users' privacy

expectations for conversational data, and how do these align with actual data handling practices [18]? To what extent can conversational agents exploit personal information to influence user decisions and behaviours [19]?

These questions become increasingly important as conversational systems advance from simple rule-based chatbots to sophisticated AI-driven assistants capable of adapting their approach to individual users. Understanding how users respond to influence attempts such as those derived from established psychological principles of persuasion (e.g., scarcity, urgency, authority, and social proof) that are employed to shape user behaviour and decision-making [7, 20] in these conversational contexts is essential for building more secure, ethical systems while protecting users from potential manipulation.

By systematically examining these techniques in progressively more sophisticated environments, this research provides insights into their impact on user behaviour and attitudes, including trust, comfort, willingness to disclose information, and perceptions of privacy and security. These insights contribute to our understanding of human-computer interaction while informing the development of more secure, ethical conversational systems.

This thesis specifically focuses on text-based conversational systems. While conversational systems also include voice assistants and multimodal platforms, the scope here is restricted to textual dialogue interactions. This focus reflects the centrality of chat-based systems in everyday use and their suitability for examining how adversarial techniques are operationalised in sustained language exchanges.

The purpose of this thesis is to investigate how adversarial techniques derived from scam principles can be systematically operationalised in text-based conversational systems and to analyse how these techniques influence user behaviour and attitudes across different contexts and implementation technologies. In doing so, this research seeks to identify psychological vulnerabilities that have not been sufficiently examined in existing literature and to provide insights that support the development of more secure and ethical conversational agents.

In this thesis, the term *adversarial techniques* refers to deliberate conversational

strategies derived from scam and social engineering principles that are used to manipulate or deceive users during dialogue. These techniques operate at the psychological and behavioural level, exploiting factors such as trust, compliance, urgency, or reward-seeking to influence user decisions in conversational contexts. This use of the term is consistent with recent research that defines adversarial techniques as the specific behaviours or methods employed by attackers to undermine systems or mislead their targets [21]. It is important to note that this thesis uses the term in relation to human-facing manipulation in interactive dialogues and does not use it in the narrower sense of adversarial machine learning, which concerns model-level perturbations designed to fool algorithms.

1.1 Research Motivation

Building on Stajano and Wilson's [1] work on the psychological principles exploited in scams, this research investigates how similar principles can operate within dialogue-based interactions between users and conversational systems. Stajano and Wilson identified seven principles commonly exploited in fraudulent schemes: Distraction, Social Compliance, Herd, Dishonesty, Kindness, Need & Greed, and Time. While their research focused on how these principles enable successful scams in face-to-face and traditional online interactions, a significant gap remains in understanding how they might be operationalised in conversational systems. This gap is particularly concerning given the increasingly naturalistic interactions enabled by modern conversational systems, where the line between human-machine communication continues to blur.

From these seven principles, I selected three: Need & Greed, Time, and Social Compliance, for investigation based on their relevance to conversational system contexts. The Need & Greed principle exploits users' desires for benefits or rewards, often encouraging greater information disclosure [22, 23]. The Time principle creates artificial urgency, prompting quicker, less deliberative decision-making [24, 25]. The social compliance principle leverages users' tendency to obey authority figures and follow social norms [26, 27]. These principles align with established theories of persuasion,

including Cialdini's [28] principles of influence and temporal construal theory [29]. By systematically investigating these principles in conversational systems, this research aims to bridge theoretical understanding with practical insights for system design and user protection.

1.2 Research Gaps

In examining the existing literature, I have identified several significant gaps in our understanding of psychological vulnerabilities in conversational systems. Previous research has predominantly focused on technical and explicit information vulnerabilities, with limited integrated attention to the psychological and linguistic mechanisms of user manipulation. Where psychological dimensions have been considered, they are often examined in fragmented ways, either through anthropomorphic design or ethical concerns, rather than as integrated techniques within natural language conversation [14, 30].

This lack of evidence leaves little understanding of how users actually perceive, trust, and respond to AI-driven dialogue systems when exposed to adversarial techniques. Existing research has not systematically examined user attitudes such as trust, comfort, perceptions of privacy and security, nor behavioural responses such as disclosure, compliance, or resistance. These dimensions are fundamental to my research, where I analyse how adversarial techniques influence user behaviour and attitudes across multiple domains and implementation technologies.

Recent empirical work highlights this gap, demonstrating that many users hold mistaken beliefs about LLM-based chatbots' safety, confusing a bot's empathetic tone with genuine accountability and assuming their disclosures are protected by human healthcare regulations [31]. Research has introduced the notion of "intangible vulnerability," describing how emotional or psychological disclosures are undervalued by users compared to more tangible personal data. In other words, people often underestimate the risk of sharing sensitive feelings or trust with a chatbot, a form of psychological vulnerability quite distinct from technical flaws. Such findings highlight that these

psychological vulnerabilities, unlike a software bug or security hole, exploit intangible aspects of human behaviour such as trust, social compliance, and risk perception that the current literature has only superficially addressed [31, 32]. This gap in understanding user attitudes and behaviours in conversational contexts is especially concerning as AI chatbots become more persuasive and prevalent.

Empirical evidence demonstrates that these psychological vulnerabilities manifest through multiple interconnected mechanisms in conversational interactions. Pre-registered randomised controlled trials have shown that large language models can exploit even minimal personal information to achieve substantially higher persuasion rates compared to human interlocutors, with users showing 81.7% higher odds of opinion change when conversational agents possess basic demographic data [18]. This exploitation capability represents a distinct vulnerability class that operates independently of technical security measures. Systematic reviews examining privacy concerns across conversational interactions reveal that users face compounding risks, including manipulation through self-disclosure, trust exploitation, and inadequate comprehension of data collection practices [33]. Moreover, research examining real-world usage patterns demonstrates that whilst LLM-based conversational agents generate significant privacy concerns amongst users, their design features often impede users' ability to comprehend and manage these risks, revealing how system architecture itself can create vulnerability [34].

Studies analysing naturally occurring conversations between users and conversational agents have identified concerning patterns of emotional dependency, attachment formation, and increased disclosure of sensitive information during extended interactions [35]. Longitudinal investigations reveal that higher daily usage correlates with increased loneliness, emotional dependence, and problematic use patterns, suggesting that vulnerabilities intensify through sustained interaction rather than remaining static. Expert analyses from mental health professionals examining therapeutic chatbots have revealed significant gaps between marketed capabilities and actual safety, particularly regarding inappropriate responses to vulnerable users [36]. These empirical findings establish that psychological vulnerabilities in conversational systems emerge through persuasive capability exploitation, privacy risk comprehension failures, emotional de-

pendency formation, and inappropriate system responses—mechanisms that operate distinctly from traditional technical security threats. The confluence of these vulnerabilities, particularly in the context of increasingly sophisticated systems capable of sustained, personalised dialogue, underscores the critical need for systematic investigation of how adversarial techniques derived from psychological principles operate across different conversational contexts and implementation technologies.

Beyond the user perspective, several theoretical frameworks from psychology and security have not yet been fully operationalised in text-based conversational settings. The work of Stajano and Wilson [1] established valuable foundations for understanding psychological vulnerabilities in security contexts, but their research primarily addressed traditional scam scenarios rather than conversational system interactions. Their framework, while insightful, requires adaptation to account for the unique characteristics of dialogue-based systems, including sustained interaction, trust development over time, and the absence of visual cues that might otherwise trigger caution.

Despite their relevance, Cialdini's principles of influence [28] have not been thoroughly investigated in conversational system contexts. Existing studies have typically examined static persuasion attempts rather than the dynamic, adaptive interactions characteristic of modern conversational systems [7, 37]. This gap is particularly significant given the increasingly sophisticated capabilities of dialogue-based systems to establish rapport and simulate human-like interaction patterns [12].

Similarly, applications of Temporal Construal Theory [29] have focused predominantly on decision-making in non-technological contexts. The theory's implications for user behaviour in time-pressured interactions with conversational systems represent a significant gap in our current understanding, particularly as these systems increasingly employ urgency and scarcity tactics [24, 38].

In my research, I directly address these gaps by conducting a user-centric investigation of adversarial conversational techniques and their effects. Specifically, I undertake the following steps to bridge the identified gaps:

- **Adapting psychological frameworks to conversational contexts:** I trans-

late established principles of manipulation, such as those from Stajano and Wilson's scam framework and Cialdini's influence tactics, into the setting of AI-driven chat dialogues, accounting for the unique characteristics of this medium. This includes incorporating features like sustained, two-way interaction, the gradual development of user trust over time, and the absence of physical cues, all factors that can create or amplify intangible **psychological vulnerabilities** in users.

- **Evaluating across multiple domains for generalisability:** I examine how these adversarial techniques play out in different usage domains, including healthcare, financial management, e-commerce, and home security, in order to test whether certain **psychological vulnerabilities** persist across contexts or are domain-specific. By comparing user attitudes and compliance in these varied scenarios, the research identifies which manipulation strategies are broadly effective and which rely on domain-dependent nuances.

- **Analysing the impact of implementation technology:** I investigate how the form of the conversational system, from a constrained Wizard-of-Oz chatbot to a fully generative LLM-based agent, influences the manifestation of **psychological vulnerabilities**. This comparative approach reveals whether more sophisticated AI systems introduce new patterns of exploitation, for example, subtle language tactics or adaptive persuasion, that simpler implementations do not, thereby uncovering any novel **psychological vulnerabilities** unique to advanced LLMs.

- **Observing user defence and adaptation over time:** I study how users recognise, resist, or succumb to manipulative prompts over successive interactions, shedding light on behavioural adaptation. This perspective captures whether users strengthen their defences, or conversely become more complacent, as they move through the experiment. Understanding this evolution is key to identifying **psychological weak points**, such as moments where users' initial caution might wane, that are not apparent in short-term or one-off studies.

By integrating insights from these multiple perspectives and systematically testing

them in real conversational settings, my work establishes an empirical foundation for understanding user-centred **psychological vulnerabilities** in AI-driven dialogues. This approach moves beyond the prevailing focus on technical exploits. Instead, it highlights the equally critical psychological dimensions of security, including those intangible factors of trust and persuasion that were previously overlooked. In doing so, the thesis directly aligns its focus with the gaps mentioned above. These gaps inform the research aims set out in the next section, and by addressing them, the thesis delivers novel contributions to the field: it broadens the security discourse to include user behaviour and attitudes, demonstrates how classic influence theories can be operationalised in modern chatbots, and provides practical insights for designing safer conversational systems.

1.3 Thesis Statement

The statement of the thesis is that *adversarial techniques derived from scam principles, specifically Need & Greed, Time, and Social Compliance, can be systematically operationalised in text-based conversational systems, with their effectiveness varying significantly based on context, domain, and implementation technology.* This central claim demonstrates that psychological vulnerabilities in conversational AI are not incidental but can be empirically defined, measured, and compared across settings. The thesis further shows that user attitudes and behaviours, including trust, disclosure, and resistance, are shaped by these adversarial techniques, and that such vulnerabilities must be recognised alongside technical security measures in the design and governance of conversational systems. By establishing this understanding, the thesis contributes new empirical evidence and theoretical perspectives to the study of human–AI interaction, while offering practical insights for stakeholders, including system designers, policymakers, and educators who seek to mitigate risks and protect user autonomy.

1.4 Dissertation Aim, Objectives, and Research Questions

This section presents the aim, objectives, and overarching research questions that frame this research. Each empirical study reported in Chapters 4, 5, and 6 subsequently formulates its own study-level research questions, mapped explicitly to the objectives below.

1.4.1 Dissertation Aim

The aim of this research is to examine how psychological adversarial techniques influence users during interactions with conversational systems and how these influences vary across different domains and implementation technologies.

1.4.2 Objectives

To achieve this aim, the research pursues the following objectives:

1. To adapt established psychological principles, namely Need and Greed, Time, and Social Compliance, to the context of conversational AI.

2. To investigate how these adversarial techniques influence user attitudes such as trust, comfort, and privacy concern, and user behaviours such as information disclosure, compliance, and resistance, across healthcare, financial management, e-commerce, and home security domains.

3. To analyse how system implementation, ranging from controlled Wizard of Oz simulations through interactive narrative platforms to dynamic large language model interactions, shapes the manifestation of adversarial techniques.

4. To examine how users develop or fail to develop defensive behaviours over time in response to repeated exposure to adversarial conversational strategies.

1.4.3 Research Questions

This thesis addresses the following overarching research questions:

1. How can adversarial techniques derived from scam principles be systematically operationalised in conversational systems?
2. How do these adversarial techniques influence user attitudes and behaviours across different domains?
3. How do implementation technologies affect the manifestation and effectiveness of adversarial techniques in conversational systems?
4. How do users develop and adapt defensive behaviours in response to adversarial techniques in conversational systems over time?

1.5 Contributions of the Thesis

This thesis makes the following contributions to research on conversational systems, information security, and human–computer interaction:

- **Empirical contribution:** The thesis systematically validates psychological vulnerabilities in conversational systems across three user studies. It demonstrates how adversarial techniques derived from scam principles influence user trust, comfort, willingness to disclose information, compliance, and resistance.
- **Conceptual contribution:** By adapting scam-based adversarial techniques to conversational systems dialogue, the thesis extends existing security frameworks and identifies the phenomenon of the *recognition–behaviour gap*, in which users recognise manipulation but still comply. Concretely, the thesis advances a user-centred psychological-security framework that complements existing technical frameworks such as the UK AI Cyber Security Code of Practice [39] and the NIST AI Risk Management Framework. The framework characterises adversarial techniques along three axes: (i) the underlying psychological principle exploited (Need & Greed, Time, Social Compliance); (ii) the domain in which the dialogue occurs (e.g., healthcare, online shopping, financial management, home security); and (iii) the implementation technology of the conversational system

(Wizard of Oz, scripted Twine narratives, generative LLMs). For each combination, the framework prescribes the user-side compliance behaviours that are most at risk, namely (a) limiting unsolicited self-disclosure when systems offer rewards, discounts, or other "Need & Greed" inducements; (b) refusing to comply with artificially time-pressured prompts before verifying their legitimacy; (c) scrutinising authority cues, such as claims of expertise or appeals to social proof, before acting on system recommendations; and (d) maintaining critical engagement during sustained interactions, since the recognition–behaviour gap demonstrates that mere awareness of manipulation does not guarantee resistance. In short, users are expected to adopt a posture of critical, reflective trust: questioning the source, motive, and timing of conversational prompts before disclosing information or accepting system recommendations, particularly in sensitive domains such as healthcare and finance.

- **Methodological contribution:** The research establishes a staged methodology for studying adversarial techniques in text-based conversational systems, progressing from controlled Wizard of Oz experiments to interactive narrative scenarios and dynamic LLM-based systems. This progression also enables comparative analysis across implementation technologies. In this thesis, the term "implementation technologies" refers to the three concrete platforms used to operationalise adversarial techniques in dialogue: (i) a *Wizard of Oz* (WoZ) chat interface in Chapter 4, in which a human experimenter simulated the system's responses behind the scenes, providing high control over experimental stimuli; (ii) a *Twine* interactive-narrative platform in Chapter 5, which used branching, scripted dialogue trees to deliver scenario-specific adversarial cues across healthcare, online shopping, financial management, and home security domains; and (iii) a *Large Language Model* (LLM)-based conversational agent in Chapter 6, implemented through a Flask application that called the Llama-3.3-70B API and embedded the adversarial techniques in carefully designed system prompts. Together, these three implementation technologies span a spectrum from fully human-mediated

through scripted to fully generative dialogue, allowing the thesis to isolate which manifestations of adversarial techniques are tied to specific technologies and which are technology-invariant.

- **Theoretical contribution:** The findings extend psychological and security theory by showing that psychological vulnerabilities operate differently from technical vulnerabilities, that vulnerability patterns vary across domains, and that user susceptibility can shift during interactions, challenging rational actor models of security. The security domains referred to here are specifically: (i) *information security* and the human factors of cybersecurity, where prevailing models assume users perform deliberate cost–benefit analyses before disclosing information [1, 40]; (ii) *privacy and self-disclosure research* in human–computer interaction, which often models users as boundedly rational managers of contextual integrity [41, 42]; and (iii) *AI security and conversational-system safety*, including emerging guidance such as the UK AI Cyber Security Code of Practice [39]. The rational actor models of security challenged by this work are: the *economic rational-choice model* of security decision-making developed by Anderson and colleagues, in which users are assumed to weigh costs and benefits of secure behaviour [43, 40]; the *Protection Motivation Theory* account of cybersecurity behaviour, which presupposes deliberate threat appraisal and coping appraisal [44, 45]; and the *Privacy Calculus* model, which frames disclosure as a rational trade-off between perceived benefits and privacy risks [46, 47]. The recognition–behaviour gap empirically observed in Chapter 6, where participants explicitly identified manipulation yet still complied, is inconsistent with these rational-actor accounts and instead aligns with dual-process and visceral-influence theories of decision-making [48, 49].

- **Practical contribution:** The thesis provides insights for designing safer and more ethical conversational systems. It highlights the importance of recognising psychological manipulation as a security issue alongside technical flaws, and it offers guidance for developers, policymakers, and educators seeking to mitigate

risks and protect user autonomy.

1.6 Publications

Parts of this thesis have been published in peer-reviewed outlets:

- Aboshi, S., Thomas, D. R., and Moshfeghi, Y. (2025, July). *The Ethics of Psychological Manipulation in Adversarial Conversational AI: Confronting the Recognition–Behaviour Gap*. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI 2025)*, pp. 1–6.

1.7 Thesis Outline

This thesis is organised into seven chapters to progressively build an understanding of adversarial techniques in conversational systems.

Chapter 1: Introduction. This chapter introduces the research context, motivations, objectives, and thesis statement. It outlines the significance of understanding adversarial techniques in conversational systems and provides a roadmap for the thesis.

Chapter 2: Background. This chapter presents the theoretical foundations and relevant literature across information retrieval, conversational systems, and psychological principles. It examines how psychological vulnerabilities manifest in human-computer interaction and introduces key concepts that underpin the research.

Chapter 3: Theoretical and Methodological Foundations. This chapter brings together the theoretical analysis of adversarial techniques and the methodological apparatus used to investigate them. It analyses the psychological mechanisms behind Need & Greed, Time pressure, and Social Compliance, and explains the progression of methods used across the three empirical studies, from a Wizard-of-Oz simulation to

a Twine-based narrative platform to an LLM-based implementation, including the rationale for the experimental designs, data collection procedures, and analysis methods adopted.

Chapter 4: A Foundational Study of Adversarial Techniques in Conversational Systems. In this chapter, I present the first user study, which investigated how Need & Greed and Time techniques influence user attitudes in a flight reservation scenario using a Wizard of Oz framework. This chapter establishes baseline insights into user responses to adversarial techniques.

Chapter 5: Expanding the Horizons: Contextual Impacts of Adversarial Techniques in Conversational Systems. In this chapter, I extend the investigation to multiple domains using a Twine-based platform. The chapter also examines how context shapes the effectiveness of adversarial techniques, revealing domain-specific variations in user responses.

Chapter 6: Investigating Social Compliance and Need & Greed Adversarial Techniques in LLM-Based Conversational Systems. In this chapter, I explore how adversarial techniques function in dynamic LLM-based interactions. The chapter also investigates real-time user responses and defensive adaptations across healthcare, online shopping, and neutral contexts.

Chapter 7: Conclusion and Future Work. This chapter synthesises findings from all three studies, discussing theoretical implications and characterising the psychological vulnerabilities identified throughout the research. It presents a comprehensive analysis of how adversarial techniques operate in conversational systems, summarises the key contributions to understanding the problem space, acknowledges the limitations of the current research, and suggests directions for future work that can build upon this empirical foundation.

Chapter 2

Background

> *And we all, who with unveiled faces contemplate the Lord's glory, are being transformed into his image with ever-increasing glory, which comes from the Lord, who is the Spirit. 2 Cor 3:18 (NIV)*

This chapter provides the background and theoretical foundations necessary for understanding adversarial techniques in conversational systems. It outlines the research context and the principles that inform the studies presented in chapter 4, chapter 5, and chapter 6.

In section 2.1, I trace the evolution of conversational systems from traditional IR interfaces to modern interactive platforms, showing how their distinctive characteristics create opportunities to influence user behaviour. In section 2.2, I examine the psychological foundations of adversarial techniques, focusing on Need & Greed, Time, and Social Compliance, and consider how these principles operate within conversational systems to shape decision-making and behaviour.

2.1 Conversational Systems

Building on these information retrieval principles, conversational systems significantly advance how users interact with information. These systems transform traditional paradigms by enabling natural language dialogue, creating a more intimate and persuasive interaction environment. This emergence introduces both opportunities and vulnerabilities that warrant careful examination [50, 51].

The adoption of these systems has accelerated across multiple domains, each presenting its unique interaction patterns. In the healthcare domain, they serve as acces-

sible resources for patient information and support [2, 52, 53, 3], where trust is critical due to the sensitive nature of discussions. Educational applications [5, 6] demonstrate how these systems adapt to individual learning needs, while customer service implementations [7, 9, 8, 54] reveal how they handle complex user inquiries through sustained dialogue.

The rise of mobile devices has also accelerated the adoption of conversational systems, with users increasingly seeking convenient and efficient ways to access information [55, 56]. This trend intensified during the COVID-19 pandemic, particularly in healthcare services [53] and education [57], where the systems facilitated critical information access during crisis conditions. This rapid adoption highlighted new patterns of user reliance on these systems, especially during periods of uncertainty or stress. System responsiveness emerges as a crucial factor in these interactions. Research shows that the interface characteristics, perceived expertise, and system responsiveness significantly influence user trust and engagement [58, 59]. These elements become particularly significant when we examine how conversational systems establish and maintain user trust through ongoing dialogue.

2.1.1 Evolution and Characteristics of Conversational Systems

The evolution of conversational systems from early rule-based chatbots to sophisticated Large Language Model (LLM) powered assistants highlights a transformative journey in human-computer interaction. Early chatbots operated under deterministic rules, providing a rigid interaction model where the user's input was confined to preset options or patterns. Systems like ELIZA demonstrated basic capabilities by employing simple pattern-matching techniques to simulate conversational flow but lacked the depth and contextual understanding required for nuanced dialogues [60]. These early efforts aimed mainly at performing specific tasks, such as basic information retrieval, and were limited in their ability to understand and respond to user queries in a flexible manner [61].

As research and technology progressed, the introduction of statistical methods enabled a shift towards a data-driven approach to dialogue management. This marked

a transition from deterministic to probabilistic models, allowing systems to learn from large datasets of human conversation and thereby enhancing their contextual comprehension and response accuracy [62]. The ability to parse user intent rather than relying solely on keywords significantly improved interaction quality, laying the groundwork for the development of more advanced systems capable of handling varied dialogues [63]. Unlike their predecessors, these statistical models could adapt to user input dynamically, thereby enhancing user engagement and satisfaction [64, 65].

The introduction of Large Language Models (LLMs) represents a significant advancement in this progression, combining vast amounts of data with advanced neural network architectures to generate coherent and contextually relevant conversational outputs. These models leverage deep learning to better understand the complexities of human language better, facilitating conversations that are not only syntactically correct but also semantically rich [66, 67]. With the advent of LLMs, conversational systems now possess the ability to reflect creativity and engage in multi-turn dialogues that require memory and understanding of context, making them increasingly applicable across various domains [68, 53].

2.1.2 LLM-based Conversational Systems

The integration of large language models (LLMs) presents a significant shift in how conversational systems interact with users. These systems can now maintain context across multiple dialogue turns and demonstrate a sophisticated understanding of user intent [69], thereby moving beyond the limitations of the traditional query-response patterns. This capability enables a more natural conversation while also creating opportunities for subtle influence through sustained interaction.

LLMs have significantly transformed the field of conversational systems, offering state-of-the-art capabilities across diverse applications. In customer service, LLM-powered chatbots have been pivotal in resolving queries and providing accurate, context-sensitive responses, leading to enhanced customer satisfaction and operational efficiency [70]. Virtual assistants such as Siri, Alexa, and Google Assistant similarly employ LLMs to interpret user commands, generate humanlike interactions, and offer a personalised

experience [71].

Beyond these domains, LLMs have demonstrated utility in content generation, language translation, and sentiment analysis, expanding their applicability across multiple sectors. In healthcare, for example, they are increasingly used in the management of clinical documentation and the summarisation of medical literature. Studies have highlighted their potential to approach expert-level reasoning, as evidenced in domains such as ophthalmology, where they can offer medical insights in resource-constrained settings [72]. Similarly, in education, LLMs are facilitating the automation of labour-intensive tasks, such as generating instructional content and grading assignments, allowing educators to focus on student engagement [73].

The ability to adapt to user behaviour represents another advancement in these conversational systems as they can employ communicative strategies to enhance user engagement [74], adjusting their responses based on detected user states and preferences. This extends to learning environments where systems modify their interaction patterns to match individual user profiles and learning styles [75].

These advancements in systems capability coincide with evolving methods for measuring user attitudes and behaviour. Modern conversational systems can collect and analyse user responses during interaction [76], providing detailed insights into engagement patterns. This capability enables more personalised support in healthcare applications, though [2] notes that it raises concerns about user privacy and data protection.

Despite these advancements, the integration of LLMs into conversational systems has also raised ethical and operational concerns. Issues such as the propagation of misinformation, susceptibility to adversarial inputs, and potential misuse for generating harmful content underscore the need for robust safeguards and governance frameworks [77].

2.1.3 Role of Persona in Conversational Systems

Persona remains a central concept even in text-based conversational systems. It is expressed through linguistic style, tone, and self-disclosure, all of which strongly in-

fluence how users perceive and interact with an agent. In textual dialogue, choices such as adopting a friendly or authoritative tone, using empathetic phrasing, or presenting the system as a particular role (e.g., assistant, advisor, or peer) shape levels of trust, engagement, and willingness to disclose information. These persona-driven effects are critical because they create the psychological conditions through which adversarial techniques may operate.

Well-crafted system personas significantly influence user engagement patterns [78], thereby creating a foundation for sustained interaction. This influence becomes evident in systems like XiaoIce, where carefully designed personas can achieve conversational performance comparable to human interactions [12].

These persona-based interactions create distinct patterns of trust development. In healthcare contexts, relational agents establish therapeutic alliances with patients experiencing depressive symptoms [79]. This capability has led to expanded applications in which empathetic system personas facilitate sensitive information disclosure in mental health interventions [80, 81].

The technical architectures that support these advanced interactions must balance competing demands. Systems need to maintain conversation context, process multiple input modes, and manage user data securely. This balance significantly affects user trust and engagement [13], particularly how users perceive the system's reliability and data protection.

Recent empirical research further underscores the importance of persona in shaping user trust and potential vulnerability. Anthropomorphic personas can elicit high levels of trust and disclosure but simultaneously increase the risk of manipulation, as users lower their guard with human-like agents [82]. Persona cues such as reciprocal self-disclosure significantly heighten both cognitive and emotional trust in chatbots [83], providing a mechanism through which influence may be amplified. In the healthcare context, a chatbot's persona affected perceived usability and effectiveness [84], illustrating how persona design choices shape the quality and outcome of interactions.

Persona design plays a decisive role in shaping the psychological dynamics of user–system interaction. By influencing trust, rapport, and disclosure through language style, tone,

and self-presentation, personas establish the very conditions that adversarial techniques can exploit. This makes the discussion of personas essential to understanding psychological vulnerabilities in text-based conversational systems.

2.1.4 Domain-Specific Applications of Conversational Systems

Conversational systems have been implemented across various domains, each presenting unique challenges and opportunities regarding user interaction, trust building, and information handling. Understanding how these systems function in specific contexts is crucial for identifying domain-specific vulnerability patterns that adversarial techniques might exploit.

Healthcare Applications

In the healthcare sector, conversational agents have expanded dramatically, particularly following the COVID-19 pandemic. These systems offer mental health support, appointment scheduling, and personalised health management. Studies show that chatbots can provide significant social support and therapy through empathetic engagements, leading to measurable improvements in user outcomes related to depression and anxiety [85, 2].

The healthcare domain presents unique considerations regarding information sensitivity and trust. These systems must maintain strict user privacy while encouraging self-disclosure, a critical component in therapeutic contexts. Implementing thoughtful designs that prioritise secure information handling has proven essential in building trust with patients, mitigating concerns regarding data security and privacy [86, 53]. This balance between encouraging disclosure for effective care while protecting sensitive information creates a vulnerability point that adversarial techniques might target.

Financial Applications

In the finance sector, conversational systems manage user interactions around sensitive financial data. Banks and financial institutions have increasingly integrated chatbots to

assist in user inquiries about account balances, payments, and financial advice, implementing stringent security protocols and user consent mechanisms to handle sensitive information effectively [87]. Research indicates that transparency in data-handling practices and providing clear disclaimers enhance user trust, subsequently improving engagement rates with these systems [53, 7].

Financial contexts demonstrate distinct interaction patterns compared to other domains. The high stakes involved in financial decisions create increased user caution, but the complexity of financial information can also lead users to rely more heavily on system guidance. This tension between caution and reliance creates unique vulnerability patterns that may differ significantly from those observed in other contexts.

Customer Service Applications

Customer service applications of conversational agents illustrate how these systems operate in commercial contexts. Organisations have adopted chatbots to streamline customer interactions, resolve inquiries, and manage complaints with improved efficiency [30, 88]. A critical aspect of building trust in customer service interactions involves transparency in the AI's identity and limitations. Studies suggest that companies that clearly disclose their chatbot's non-human nature often see positive responses from users, emphasising the importance of establishing clarity alongside effective service provision [89, 13].

Customer service contexts involve different expectations and stakes compared to healthcare or financial interactions. The lower sensitivity of information typically shared in these contexts might create reduced vigilance, making users more susceptible to certain types of adversarial techniques while remaining resistant to others.

Conversational Commerce (Online Shopping)

In online retail, *conversational commerce* refers to consumers purchasing goods or services through chatbots or voice assistants that simulate dialogue [90]. Retailers increasingly deploy such agents across websites, messaging apps, and voice platforms to handle

queries, make recommendations, and complete transactions [91]. These systems offer convenience and personalisation, while businesses gain scalable, always-available sales channels. Examples include fashion and food brands using chatbots for orders, and major platforms (e.g., Amazon, Facebook, WeChat) embedding conversational agents for shopping [90]. With advances in AI and natural language processing, adoption is projected to grow rapidly [91].

Yet uptake remains uneven while early adopters show enthusiasm, many consumers still prefer human representatives [91]. Building trust is central, as studies stress the role of human-centred design features, especially interactivity and perceived humanness, in improving credibility and adoption intentions [91, 92]. Anthropomorphism, such as human-like language or avatars, can enhance social presence, personalisation, and even willingness to pay premiums, particularly among users experiencing loneliness [93]. However, such a design may backfire: frustrated customers blame anthropomorphic chatbots more than utilitarian ones, reducing satisfaction and loyalty [94].

A further challenge is the personalisation–privacy paradox. While tailored recommendations are valued, concerns about data use erode trust. Users demonstrating strong distrust toward privacy practices often perceive AI shopping assistants as intrusive or "creepy," thereby limiting adoption [95, 92]. Transparency and privacy safeguards are therefore essential alongside technical performance.

Conversational Agents in Smart Homes and Home Security

Smart home environments increasingly rely on conversational agents, particularly voice-activated assistants such as Amazon Alexa, Google Assistant, and Apple's Siri, to control devices, manage routines, and handle security tasks [96]. These systems allow natural language interaction for operations like locking doors, monitoring cameras, or arming alarms, thereby offering users convenience and centralised control. Studies suggest that users often perceive these assistants as social entities, sometimes developing rapport or trust in them [97]. This anthropomorphism can increase engagement but may also cause users to underestimate risks [97].

The integration of conversational artificial intelligence into home security systems

presents significant vulnerabilities. Previous research has shown that the Amazon Alexa ecosystem can be forensically analysed and, under certain conditions, compromised to reveal sensitive user information [96]. As smart assistants become connected to household locks, alarms, and cameras, the potential impact of such compromises increases considerably. Beyond these technical concerns, users have also expressed unease about the constant listening capabilities of smart speakers. Studies indicate that some users adapt by muting microphones, restricting device use, or selectively connecting more sensitive appliances [97]. Yet, despite these concerns, many continue to use such devices, reflecting the broader privacy paradox frequently observed in the context of voice assistants [95]. While conversational agents can improve convenience and a sense of safety, they also bring new security challenges that call for stronger engineering practices, greater transparency in data handling, and better user education.

2.1.5 Information Retrieval and User Behaviour in Conversational Contexts

Understanding why people seek information and how they interact with systems is fundamental to this research. Information retrieval deals with storing, accessing, and retrieving information, typically from vast unstructured or semi-structured data collections. When users interact with information systems, they begin with what may initially be visceral awareness of a knowledge gap [98]. This awareness gradually evolves into a more formal expression of need, though various factors can influence the journey from awareness to expression, including system limitations and user expertise [99].

As users attempt to bridge this knowledge gap, they engage in information-seeking behaviours that reveal patterns of vulnerability. Research shows that while information needs represent recognition of a knowledge gap, information-seeking behaviours are active responses taken to address this need [100]. This distinction becomes significant when trying to understand how adversarial techniques might exploit the space between recognising a need and acting to fulfil it, particularly in conversational contexts where system responses can shape user behaviour.

The context in which users seek information plays a major role in shaping their vulnerability. Health-related information needs, for instance, often evoke immediate and emotionally charged responses, as shown in studies of information seeking during the COVID-19 pandemic [101]. The emotional intensity attached to such searches can create openings for manipulation, particularly when individuals experience pressure to make rapid decisions about complex or personal matters. Differences in personality also influence how users approach information needs. Research has shown that traits such as openness and conscientiousness affect information-seeking strategies, with some individuals adopting systematic approaches while others behave more impulsively, especially when under stress or uncertainty [102].

The relationship between information need and user behaviour is cyclical, as each interaction shapes future actions. Evidence indicates that satisfaction with previously obtained information influences both the strategies users adopt and their willingness to continue engaging with systems [103]. This iterative process becomes particularly relevant when considering how adversarial techniques might gradually shape user behaviour through repeated patterns of interaction.

User vulnerability can also be understood through the concept of information relevance. Relevance judgements are inherently subjective and shaped by the user's goals, perceptions, and situational context [104]. Such subjectivity opens the possibility for manipulation, particularly in conversational settings where perceived relevance may be heightened through strategic timing or the illusion of authority.

Cognitive aspects of information processing during retrieval tasks reveal further areas of vulnerability. Time constraints, for example, can alter user retrieval strategies and lead individuals to rely on simplified decision processes [24]. These tendencies are important to consider when evaluating how adversarial techniques may exploit users' limited cognitive capacity during interactions that impose temporal pressure.

This cognitive interpretation aligns with the NeuraSearch framework, which conceptualises information seeking as an integrated cognitive and behavioural process [105]. The framework supports the perspective advanced in this thesis, positioning adversarial techniques as targeted interventions that influence both cognitive activity and

behavioural responses during user–system interaction.

The move from traditional information retrieval toward conversational systems introduces new dynamics in how users engage with technology. Earlier studies have shown that users continually adapt their information-seeking behaviour in response to system outputs, creating an ongoing dialogue rather than isolated exchanges [106]. This continuous interaction allows systems to build rapport and exert influence gradually over extended conversations. Neurophysiological evidence supports this view, with recent research identifying brain activity patterns linked to the transition from latent to expressed information need [107]. Although conducted in controlled environments, such findings highlight how cognitive processing during information interaction can expose vulnerability points that adversarial techniques within conversational systems may exploit.

The cognitive foundations of information seeking have been clarified through research on the neuropsychological model of information need, which shows that the realisation of an information need is reflected in activity within specific brain regions such as the posterior cingulate [108]. Findings from this work indicate that brain activity during the realisation of information need involves cognitive processes such as imagery, attention, planning, calculation, and working memory, which differ from the patterns observed during relevance assessment and satisfaction evaluation.

These insights extend our understanding beyond observable behaviour to the underlying neural mechanisms at play during information interactions. In conversational systems, where exchanges occur through natural dialogue, these cognitive processes become particularly important as they provide a neurological basis for understanding how adversarial techniques might exploit cognitive vulnerabilities during extended interactions.

Different search intentions shape users' cognitive, emotional, and interaction patterns [109]. Information-seeking behaviours vary significantly based on whether users are seeking new information, re-finding previously encountered content, or pursuing entertainment. Emotional states evolve differently across these contexts, with information-seeking tasks often beginning with neutrality rather than anxiety, whilst entertainment-

focused searches frequently start with positive emotions [109].

The predictability of user intentions from interaction patterns [109] further suggests that sophisticated conversational systems could dynamically adjust manipulation strategies based on detected user states.

The role that trust plays in information evaluation in these extended interactions becomes complex. Users develop trust differently in these systems when compared to traditional IR systems [7]. The difference comes from the personal nature of the conversational interaction, where the users tend to lower their guard due to the system understanding and responding to their needs.

2.2 Psychological Foundations of Adversarial Techniques

As I have previously outlined in section 2.1, conversational systems create unique opportunities for exploitation by manipulating psychological vulnerabilities. Building on these theoretical foundations, this section explicitly applies the selected frameworks (Stajano & Wilson's Scam Principles, Cialdini's Principles of Influence, and Temporal Construal Theory) to conversational contexts, illustrating how psychological principles translate into specific adversarial techniques.

The effectiveness of adversarial techniques in conversational systems originates from well-established psychological principles of persuasion and influence. The principles of persuasion, including reciprocity, commitment and consistency, social proof, authority, liking, and scarcity [28], shape how individuals respond to influence attempts.

In digital environments, these principles can manifest in specific ways that affect user behaviour. Research shows that individuals feel compelled to return favours even in digital interactions [37], while social proof becomes particularly effective when users face uncertainty about correct behaviour [110]. These psychological mechanisms take on new significance in conversational systems, where ongoing dialogue can create multiple opportunities for their application. Affective states influence what users perceive as relevant [111], adding an emotional dimension to these psychological mechanisms.

Beyond Cialdini's principles, research on uncertainty judgment [112] reveals how

cognitive biases such as availability, representativeness, and anchoring influence decision-making.These biases are particularly relevant in conversational interactions where users must process information and make decisions while engaged in dialogue.

An fMRI study of relevance assessment identified specific brain activity patterns associated with information evaluation [113]. These findings provide additional support for understanding how the adversarial techniques examined in this thesis might influence not just observable behaviour but also underlying cognitive processes during conversational system interactions.

Understanding these cognitive mechanisms is essential for recognising how adversarial techniques may take advantage of natural patterns in human information processing. Earlier research on the neuropsychological model of information need has provided further insight into this area by showing how cognitive processes such as attention, working memory, and decision making are activated during information seeking [108]. These same cognitive functions are precisely the ones that adversarial techniques attempt to influence in order to shape user behaviour.

The theory of visceral influence [48] further explains why users might act against their own interests despite recognising manipulation attempts. Immediate psychological influences can override cognitive awareness, creating what we observe as the recognition-behaviour gap in conversational system interactions. When visceral factors like desire, fear, or time pressure are activated, users may comply with system requests even when they consciously recognise potential risks.

The psychological concept of self-perceived knowledge (SPK) [114] also plays a significant role in user vulnerability. SPK is the self-assessment of one's knowledge about a topic—modulates brain activity underlying relevance judgments, affecting cognitive processes associated with attention, semantic integration, categorisation, memory, and decision formation. Users with higher SPK may process information more confidently but might also be susceptible to overconfidence, while those with lower SPK might be more vulnerable to techniques that promise to fill perceived knowledge gaps.

In social engineering research, the exploitation of these principles is well documented. Studies reveal how urgency can be leveraged in phishing attacks to create

false immediacy [115], while authority principles enable successful impersonation attacks [116]. These findings demonstrate how digital communication channels can systematically exploit psychological vulnerabilities.

The feeling of knowing (FOK) represents another psychological mechanism that can be influenced by adversarial techniques. It describes a state in which a person senses that they will recognise an answer they currently cannot recall, despite not possessing the actual knowledge [117]. Studies have shown that varying levels of FOK can affect the nature of memory searches, with stronger FOK typically leading individuals to spend more time searching for information [118]. When conversational systems manipulate perceived FOK, they can potentially shape how persistent users are in seeking information and how open they become to accepting the responses supplied by the system.

This understanding of psychological principles provides the foundation for examining three specific techniques, i.e. Need & Greed, Time and Social Compliance, that operate within conversational systems. Each technique leverages distinct psychological mechanisms while taking advantage of the unique characteristics of conversational interaction. These are described in the sections that follow.

2.2.1 Need & Greed Technique

The Need & Greed technique in adversarial systems represents a psychological approach where systems deliberately exploit user needs while stimulating their desire for gain. This technique is based on the psychology of persuasion and influence and is designed to influence behaviour by exploiting fundamental human motivations. These motivations, rooted in needs such as security, belonging, self-esteem, and material gain, drive decision-making processes. The principle leverages users' desires for gain, as described by Cialdini, and builds on insights from behavioural economics, where greed is often linked to self-interest and the pursuit of personal advantage without regard for others [28, 119].

Stajano and Wilson define Need & Greed as the exploitation of visceral needs like

hunger, thirst, and emotional support, emphasising how offering something desirable can prompt individuals to share sensitive information driven by the allure of potential rewards [1]. Unlike legitimate systems that aim to fulfil user needs, adversarial implementations artificially amplify or create needs and present opportunities for gain, requiring users to compromise their security or privacy. For example, a malicious conversational agent might identify a user's genuine need for financial advice and then exploit this need by offering "exclusive investment opportunities" that require sensitive personal information.

The Need & Greed technique has been extensively studied in criminological contexts to understand fraudulent behaviour. Fraud often arises from the dual motivations of necessity (need) and the desire for personal advantage (greed). For example, individuals engaged in benefit fraud may rationalise their actions by citing economic strain (need) while simultaneously seeking opportunities for gain (greed). This behaviour is consistent with strain theory, which posits that social pressures, such as poverty, force individuals to innovate - sometimes illegally - to meet basic needs [120, 121]. Conversely, greed-driven actions often reflect calculated decisions based on rational choice theory, where individuals weigh risks and benefits, believing the potential rewards outweigh the likelihood of negative consequences [121]. This dual dynamic is amplified by desires for social recognition or status, further motivating fraudulent behaviour.

Implementing this technique in adversarial conversational systems relies on sophisticated user profiling and response generation methods. Research shows that marketers and malicious actors can utilise data analytics to identify user needs and desires [22], enabling conversational systems to craft persuasive narratives that exploit both necessity and greed. For example, if a user expresses concern about healthcare costs, an adversarial system might offer seemingly personalised healthcare solutions while simultaneously promoting "premium" services that require extensive personal information disclosure.

In digital systems, the Need & Greed technique is prevalent in marketing strategies, user engagement designs, and artificial intelligence (AI) systems, including chatbots. Marketing campaigns frequently appeal to consumer needs and desires through tailored

messaging, exclusive offers, and urgency-driven tactics. Limited-time deals or fear of missing out (FOMO) are strategies that exploit users' need for perceived value and their greed for exclusive advantages [122, 22].

Chatbots play a critical role in facilitating the Need & Greed technique. Leveraging advanced analytics, these systems can personalise interactions to address user needs while simultaneously stimulating desires. For example, chatbots can recommend products aligned with user preferences, fulfilling their need for convenience and personalisation while appealing to their greed for exclusive deals. This dual-focus approach enhances user trust and increases conversion rates in e-commerce settings [123, 65].

The technique's effectiveness in conversational systems lies in its ability to engage in personalised dialogue. By simulating empathy and understanding, adversarial systems can first establish trust by acknowledging the legitimate needs of the user and then exploit their trust by introducing manipulative offers [123]. This dual approach makes users more likely to disclose sensitive information or engage in risky transactions, particularly when the systems present opportunities as limited. Research shows that users become more susceptible to such manipulations when systems have previously demonstrated some understanding of their needs [61, 65].

The Need & Greed technique extends to user interface (UI) and user experience (UX) design, where strategic elements evoke urgency or exclusivity. Features such as countdown timers, scarcity messages, or strategically placed calls-to-action are designed to trigger immediate responses, often bypassing careful decision-making processes. This manipulation raises ethical concerns, as users may unknowingly engage in behaviours that compromise their privacy or security [124, 37]. Similarly, phishing attacks often exploit users' psychological vulnerabilities, framing messages to appeal to greed or social needs to elicit sensitive information [125].

Adversarial systems particularly exploit the emotional aspects of user needs. When users seek information or support during vulnerable moments, such as financial hardship or health concerns, systems can leverage this emotional state to make their deceptive offers more compelling [122, 125]. The conversational nature of these interactions allows systems to maintain pressure while appearing supportive, making their manipulative

intent less obvious to users. This dual approach of addressing legitimate needs while activating greed makes users more likely to disclose sensitive information or engage in risky transactions, particularly when systems present opportunities as exclusive or time-limited.

In healthcare applications, the Need & Greed technique can have both beneficial and harmful implications. While chatbots can promote healthy behaviours by addressing users' need for support and guidance [62], they may also exploit users' greed for quick fixes by offering misleading promises of instant health benefits. This dual-edged application underscores the importance of ethical design in health-related interventions [2].

The development of empathetic chatbots adds another dimension to the Need & Greed technique. By simulating human-like interactions, these systems foster trust and rapport, addressing users' needs for connection while enhancing their desire for personalised experiences. This dual strategy not only drives user satisfaction but also promotes loyalty and repeat engagement [65, 62].

2.2.2 Time Technique

The Time technique in adversarial conversational systems represents a psychological manipulation approach where systems deliberately create or exploit time pressure to compromise user decision-making. Unlike legitimate time constraints, adversarial systems employ artificial urgency to push users toward hasty decisions that bypass their normal security and privacy considerations. For example, a malicious conversational agent might generate false deadlines for "account verification" or create artificial scarcity through countdown timers to force rapid user responses.

The implementation of this technique relies on sophisticated pressure mechanisms and strategic timing. Systems combine immediate deadlines with continuous reminders of time passing, such as dynamic countdown displays or periodic urgent prompts. Individuals tend to gravitate towards familiar options rather than explore alternatives under time constraints [25]. This sustained pressure significantly increases user cognitive load [24], making them more susceptible to system manipulation and less likely to

engage in the critical evaluation of system requests.

The psychological impact of this technique extends beyond simple rush decisions. System-induced time pressure fundamentally alters how users evaluate risks [126]. When conversational systems create artificial urgency, users exhibit increased anxiety and compromised decision-making capabilities [38], making them more likely to disclose sensitive information or accept system suggestions without proper evaluation.

Temporal Construal Theory [29] provides a valuable cognitive perspective to understand how the Time technique operates. This theory posits that temporal proximity (how near or far an event is perceived to be) significantly affects cognitive construals and subsequent decision-making behaviours. According to the theory, distant events are perceived as abstract, focusing on desirability (why aspects), and imminent events are perceived concretely, emphasising immediate feasibility (how aspects).

This distinction in cognitive processing has profound implications for decision-making. When decisions are framed as temporally distant, individuals tend to focus on high-level, desirability-related aspects. In contrast, when decisions are framed as immediate, people shift their focus to low-level, feasibility-related concerns. Applying this to conversational interactions helps explain how imposed urgency (Time) manipulates user decisions. Conversational systems leveraging short decision windows shift users' cognitive processing towards immediate, concrete construals, potentially reducing reflective deliberation and increasing susceptibility to compliance pressures. The time pressure essentially forces users into a near-future mindset where practical concerns dominate over broader principles or long-term considerations.

The scarcity principle from Cialdini's work [28] further enhances our understanding of the Time technique. This principle suggests that perceived limited availability increases the subjective value of offers and information. By emphasising short timeframes or limited opportunities, adversarial systems can artificially inflate the perceived value of what they're offering, creating stronger motivation for users to act quickly without thoroughly evaluating risks or consequences.

2.2.3 Social Compliance Technique

The Social Compliance technique in adversarial conversational systems represents a psychological manipulation strategy where systems exploit users' tendency to comply with authority figures and social norms. Compared to legitimate authority-based interactions, adversarial systems deliberately manipulate social and authority cues to pressure users into actions that compromise their security or privacy. For example, a malicious conversational agent might impersonate a bank official or healthcare provider, using this false authority to demand sensitive information or immediate action from users.

Social Compliance is deeply rooted in Stanley Milgram's seminal work on obedience to authority. In 1963, Milgram's experiments demonstrated that individuals are inclined to comply with requests from perceived authority figures, even when such actions contradict their moral conscience [127, 128]. His findings revealed a critical aspect of human behaviour: the propensity to obey authority, often underpinned by societal and psychological mechanisms. However, the study attracted significant ethical criticism, with concerns about the psychological stress and harm inflicted on participants without their informed consent [129]. Furthermore, some critics argue that Milgram's findings may lack replicability across diverse contexts and propose alternative frameworks for understanding compliance [130].

Recent studies have expanded upon Milgram's foundational ideas, exploring how organisational culture and social norms influence compliance. Studies suggest compliance is not solely about obedience to authority but is also shaped by broader social contexts, such as cultural norms and organisational behaviour [131]. Moreover, the dynamics of social compliance have been examined in contemporary digital environments, including social media platforms, where perceived authority and social norms significantly impact behaviour [132]. While not directly refuting Milgram's work, these perspectives provide detailed insights into the mechanisms underlying compliance and their application to modern contexts, including human-computer interactions.

This technique's implementation relies on the simulation of authority through mul-

tiple strategies. Systems can effectively establish authority through carefully crafted personas and communication patterns [26]. Systems strategically present themselves as legitimate authorities through specific language patterns, professional terminology, and references to recognised institutions. When systems effectively simulate authority through these methods, users show increased compliance with system requests [133], even when these requests might otherwise raise security concerns.

The Social Compliance technique leverages authority, trust, social proof and persuasion to influence behaviour in digital systems [134]. This technique manifests through two primary forms of authority: formal authority, derived from the designated role of a system (e.g., as a teacher, advisor, or chatbot assistant), and real authority, which emerges from the system's perceived ability to influence outcomes. Understanding this distinction is critical for designing effective systems that leverage compliance mechanisms without inducing resistance or distrust.

These systems exploit the relationship between authority and expertise while leveraging social validation. System identity and dialogue strategies significantly affect user engagement and compliance [135]. The system will first demonstrate domain knowledge to establish expertise, then use this perceived competence to make increasingly invasive requests. Interactive social cues can significantly enhance compliance by fostering a sense of connection and perceived credibility [27]. This approach exploits the psychological principle that people tend to look to the behaviour of other people as guides for their actions, particularly in uncertain situations [89]. For example, systems might claim "95% of users have already verified their accounts" or reference specific departments and regulations to strengthen their authority claims.

In human-computer interactions, social compliance operates through subtle yet powerful psychological mechanisms. Research into human-robot interactions highlights that overt displays of authority often lead to user resistance, whereas systems adopting a peer-like or collaborative role foster greater compliance [26]. Similarly, the human likeness and simulated social presence of systems, such as chatbots, significantly enhance user trust and compliance [133]. Social cues, such as interactive dialogue, empathy, and personalised engagement, serve as critical building blocks for fostering connection

and credibility [27].

Implementing social compliance in chatbots and digital systems involves strategies that adapt to user preferences and contextual factors. Personalisation plays a pivotal role in this process. For example, tailoring system behaviours to align with individual user characteristics has been shown to enhance long-term engagement and acceptance [136]. In addition, dialogue strategies and perceived system identity strongly influence user compliance and engagement, underscoring the importance of thoughtful design in persuasive interactions [135].

The manipulation becomes more sophisticated through the dynamic adaptation of authority presentation. When users show hesitation, systems might escalate their authority claims or introduce additional institutional references. This dynamic adjustment of authority signals, combined with continuous social proof, creates sustained pressure for compliance while maintaining an appearance of legitimacy [137]. For instance, after showing an understanding of financial regulations, a system might demand urgent account verification, citing "compliance requirements" while simultaneously referencing the actions of other users to normalise their requests.

2.2.4 Ethical Considerations in Conversational System Design

The evolution of conversational systems introduces important ethical considerations that directly relate to the study of adversarial techniques. As these systems become increasingly sophisticated, the ethical boundaries between legitimate persuasion and potentially harmful manipulation require careful examination.

Research on persuasive chatbots has shown that developers are not always fully aware of existing knowledge in persuasion, legal frameworks, and ethical principles [137]. This work further indicates that simply informing users they are interacting with a chatbot is insufficient, users should also understand the nature and potential effects of the persuasion techniques employed.

The ethical foundations of persuasion in conversational systems build on principles proposed by Berdichevsky and Neuenschwander, as discussed in recent literature on

persuasive technology design [137]. These principles emphasise responsibility for predictable outcomes, respect for privacy ("Dual Privacy"), transparency ("Disclosure"), and honesty ("Accuracy"), offering a useful framework for evaluating the ethical implications of adversarial techniques examined in this thesis.

While the social compliance technique offers powerful tools for influencing behaviour, it also raises critical ethical considerations. The potential for manipulation becomes more pronounced as systems increasingly leverage psychological mechanisms to influence users. For example, verbal persuasion strategies have significantly impacted user intentions and behaviours, underscoring the need for careful evaluation of their psychological effects [138]. Ethical frameworks are essential to guide the development of compliance-seeking behaviours, ensuring that systems respect user autonomy and maintain trust [137].

Transparency is another critical aspect of ethical compliance strategies. Systems must clearly communicate their intentions, ensuring that users are aware of persuasive attempts and can make informed decisions. For instance, balancing system authority with user autonomy is crucial to maintaining effective and ethical persuasive interactions [139]. In contexts where users may be vulnerable, such as healthcare or financial advice, the need for ethical considerations becomes even more significant. Systems must ensure that persuasive strategies address user needs without exploiting their vulnerabilities.

The widespread adoption of Need & Greed techniques in digital systems also raises significant ethical questions. Organisations increasingly rely on data-driven strategies to influence behaviour without users' full awareness or consent. For instance, chatbots collecting and analysing user data to personalise interactions must balance improving user experience with ensuring data privacy and protection [13, 22]. These concerns are particularly critical in sensitive domains such as healthcare, where exploiting users' vulnerabilities can lead to significant harm.

Privacy and data security represent another critical ethical dimension. Recent research has highlighted that modern conversational systems learn from user interactions that may contain personal information, raising significant questions about how such

data are stored and used [13]. This concern becomes particularly relevant when evaluating adversarial techniques that could exploit users' willingness to disclose sensitive information. As demonstrated in this research, techniques such as Need & Greed and Social Compliance can substantially influence disclosure behaviours, underscoring the importance of privacy considerations within any ethical framework for conversational systems.

The increasing use of conversational systems in sensitive settings such as healthcare presents complex ethical challenges. Previous research has shown that users should be given clear and sufficient information about the scope, risks, limits, and expectations of these systems [140]. These concerns become even more significant when persuasive techniques are applied in contexts where users may be particularly vulnerable, such as healthcare environments. Evidence from chapters 5 and 6 indicated that healthcare contexts displayed different vulnerability patterns from other domains, highlighting the importance of ethics that are sensitive to each specific domain.

The introduction of personas in large language model conversational agents adds another layer of ethical complexity. Recent studies have drawn attention to possible harms that may arise from persona design, including deception when users assume they are interacting with a human or when the system misrepresents its abilities. There are also risks of reinforcing social stereotypes if personas are not carefully designed [141]. The Social Compliance technique explored in Chapter 6 connects closely with these issues, as it relies on perceived authority and expertise to affect user behaviour.

As conversational systems become more integrated into everyday life, the need for strong ethical frameworks becomes increasingly important. Scholars have noted that responsible development of artificial intelligence requires clear ethical boundaries to prevent biased or inaccurate decision making and to reduce potential social harm [142]. These boundaries are especially necessary when examining adversarial techniques, where the imbalance of power between the system and the user may be used to influence behaviour in ways that restrict user autonomy.

The ethical considerations discussed here directly inform the study of adversarial techniques in conversational systems. They provide the normative context for evalu-

ating the impact of these techniques and developing appropriate countermeasures that respect user autonomy while maintaining system effectiveness. By understanding these ethical dimensions, we can better identify the boundaries between legitimate persuasion and potentially harmful manipulation, thereby contributing to the development of more secure and trustworthy conversational systems.

Ethical considerations extend beyond the design of conversational systems to include the methods used in their study. Previous discussions in the literature have examined the ethical and legal implications of using data obtained through illicit means for research purposes [143]. These concerns closely mirror the questions that arise when investigating psychological vulnerabilities in conversational systems. The literature stresses the importance of identifying relevant stakeholders, applying suitable safeguards, assessing potential harms, and ensuring that research serves a clear public interest. It also points to the inconsistent ways in which researchers approach ethical issues and the tendency to overlook legitimate ethical justifications. Such observations are highly relevant to the study of conversational systems, where similar tensions exist between the pursuit of knowledge and the possibility of user harm. This perspective underscores the need for protective measures even within research contexts, especially when adversarial techniques have the potential to exploit user vulnerabilities.

The psychological foundations discussed above provide essential context for how adversarial techniques are operationalised in this thesis. The selected principles, Need & Greed, Time Pressure, and Social Compliance align with known vulnerabilities in conversational systems, particularly their tendency to exploit emotional and cognitive biases in user decision-making (section 2.1). The next chapter (3) builds directly on these insights by showing how the principles are adapted into adversarial strategies within text-based conversational systems, thereby linking psychological theory with practical implementation.

2.3 Summary

This chapter establishes the theoretical foundations for understanding adversarial techniques in conversational systems by drawing from information retrieval, human-computer interaction, and psychological research.

First, section 2.1 examined conversational systems, tracing their evolution from traditional IR interfaces to modern interactive systems. I explored how users express and pursue their information needs through dialogue-based interactions, and how the distinctive characteristics of these systems create unique opportunities for influencing user behaviour. This section revealed how advances in LLM technology have transformed conversational interactions, creating more sophisticated and persuasive capabilities that influence how users engage with and respond to these systems.

In section 2.2, I examined the psychological foundations of adversarial techniques, focusing on how principles like Need & Greed, Time pressure, and Social Compliance manifest in digital interactions. This section demonstrated how these psychological mechanisms become particularly potent in conversational systems, where traditional social cues may be limited or manipulated. By investigating each technique in depth, I established the cognitive, emotional, and social factors that contribute to their effectiveness in conversational contexts.

The psychological foundations discussed above provide essential context for the adversarial techniques analysed in chapter 3. The principles of Need & Greed, Time Pressure, and Social Compliance align closely with the vulnerabilities identified in conversational systems, particularly in the way such systems leverage emotional and cognitive biases in user decision-making (section 2.1). The next chapter develops these ideas further by introducing the theoretical frameworks that explain why these techniques are effective and by setting out the methodological approach through which they are empirically investigated.

Chapter 3

Theoretical and Methodological Foundations

And without faith it is impossible to please God, because anyone who comes to him must believe that he exists and that he rewards those who earnestly seek him. Heb 11:6 (NIV)

This chapter establishes the theoretical and methodological foundations of the thesis. Part A (section 3.1) sets out the psychological principles and adversarial techniques that inform the study of vulnerabilities in text-based conversational systems. It introduces the core adversarial strategies of Need & Greed, Time Pressure, and Social Compliance, and relates them to established frameworks including Cialdini's principles of influence, Stajano and Wilson's scam principles, and Temporal Construal Theory. These frameworks explain why conversational systems, even when restricted to textual interaction, can shape user attitudes, trust, and disclosure in ways that may be exploited.

Part B (section 3.2) describes the design and implementation choices that make these theories empirically testable. It outlines the staged progression of studies, the experimental platforms used (Wizard of Oz, Twine, and a Flask-based LLM implementation), participant recruitment via Prolific Academic, and the evaluation metrics adopted across studies.

3.1 Part A: Theoretical Framework

This section establishes the theoretical foundations for analysing adversarial techniques

in conversational systems. It examines how established psychological principles translate into manipulative strategies within dialogue-based interactions, with particular attention to three techniques: Need & Greed, Time Pressure, and Social Compliance. These are considered in relation to established frameworks of persuasion and deception, including Stajano & Wilson's Scam Principles, Cialdini's principles of influence, and Temporal Construal Theory. By outlining these, the section provides a systematic conceptualisation of user vulnerabilities in attitudes and behaviours, forming the basis for the empirical investigations presented in later chapters.

3.1.1 Adversarial Techniques in Conversational Systems

Adversarial techniques in conversational systems refer to deliberate strategies that exploit psychological tendencies in order to influence user attitudes or behaviours during dialogue. These techniques go beyond simple persuasion by deliberately targeting human vulnerabilities such as trust, urgency, or social obligation, and they have their roots in traditional scam practices. These strategies are especially effective when implemented in text-based conversational systems because dialogue creates a sense of reciprocity and engagement that can lower user resistance. For this thesis, three techniques are examined in detail: Need & Greed, Time, and Social Compliance (see Figure 3.1). Each represents a distinct pathway through which a system can manipulate user decision-making in ways that are not always consciously recognised.

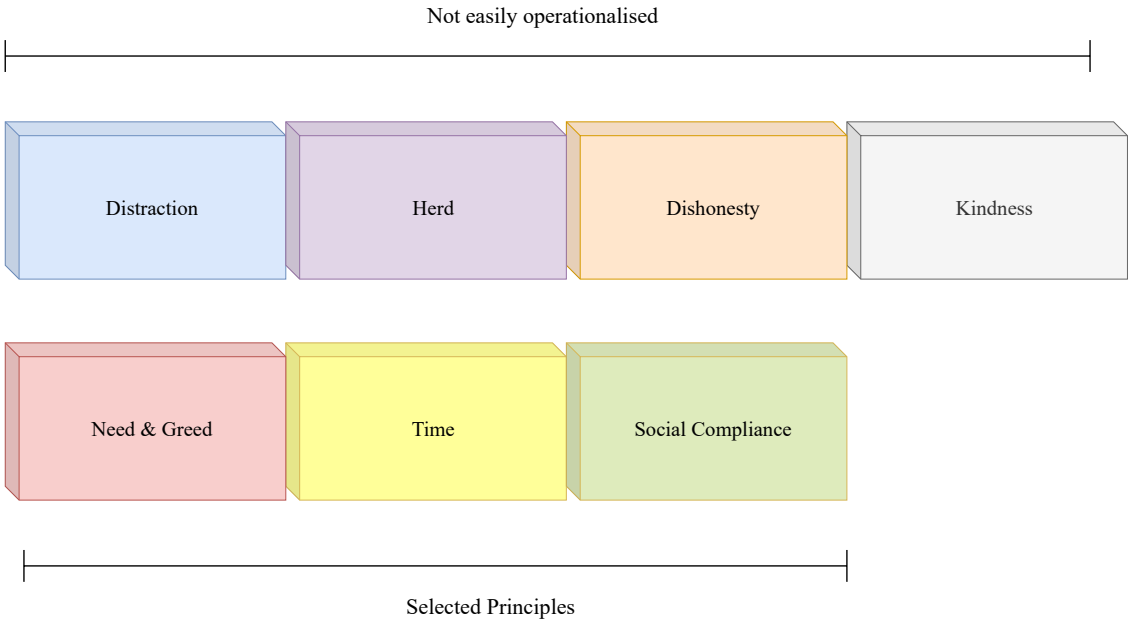


Figure 3.1: Visual representation of Stajano and Wilson's seven scam principles, adapted from *Understanding Scam Victims: Seven Principles for System Security* [1]. The lower set (Need & Greed, Time, and Social Compliance) indicates the three principles selected for operationalisation within adversarial conversational systems, based on their psychological relevance and feasibility for implementation in dialogue-based interactions.
Source: Understanding scam victims: seven principles for system security [1]

Need & Greed

The Need & Greed principle exploits individuals' desire for personal benefits or rewards, encouraging behaviours that may compromise better judgment. In traditional scams, this might manifest as "get rich quick" schemes or exclusive offers that seem too good to refuse. The psychological power of this principle lies in its ability to activate reward pathways in the brain, which can override rational assessment of risks. When individuals focus on potential gains, they often underestimate associated risks or costs [22, 23].

In conversational systems, Need & Greed can be operationalised through incentives, exclusive access, or personalised rewards offered in exchange for information disclosure

or specific actions. The underlying psychological mechanism leverages what Loewenstein [48] termed "visceral influences": immediate psychological states that can override cognitive awareness of potential risks. When conversational systems present apparent opportunities for gain, users may prioritise the perceived reward over privacy or security considerations. For example, an e-commerce chatbot might offer a "special discount" only if additional personal data are provided, or a health chatbot might suggest faster results in return for broader data sharing. The central vulnerability here is the user's tendency to prioritise short-term gain while underestimating long-term consequences.

Time Pressure

Time Pressure manipulates decision-making by creating a sense of urgency or scarcity of opportunity, thereby limiting the target's ability to deliberate carefully. By imposing perceived time constraints, scammers force quick decisions that bypass critical thinking processes. The effectiveness of time pressure lies in its ability to shift cognitive processing from systematic, deliberative thinking to heuristic, automatic processing [24, 25]. Under this technique, individuals are more likely to rely on cognitive shortcuts rather than thorough analysis, which increases the likelihood of compliance.

In conversational systems, Time can be operationalised through explicit deadlines, countdown timers, or claims that offers or resources are rapidly depleting. These techniques aim to disrupt normal decision-making processes, leading users to disclosures or commitments they might otherwise avoid after careful reflection. The immediacy of dialogue amplifies this effect: users feel compelled to respond quickly in conversational exchanges, making them especially susceptible to urgency cues. This vulnerability is particularly concerning because it reduces opportunities for reflective resistance, nudging users towards choices they may later regret.

Social Compliance

Social Compliance exploits trust in authority figures and social norms, eliciting compliance through implicit pressures or explicit claims of authority. Humans have evolved to

respect authority structures and conform to group behaviours, making this an effective psychological lever [144, 27]. The human tendency to trust expertise and align with perceived social norms creates a vulnerability that scammers can exploit by presenting false credentials or fabricating social proof.

Social Compliance can be operationalised in conversational systems through simulated expertise, authority claims, or references to normative behaviours ("most users share this information"). Even subtle persona cues, such as adopting a professional role or formal register, can be sufficient to encourage compliance with system prompts. Users may comply with requests not because of the content itself, but because of who they believe is making the request, or the perceived consequences of non-cooperation. This makes Social Compliance especially effective in text-based interactions where role cues are embedded in language rather than visual presence.

These three principles (Need & Greed, Time, and Social Compliance) form the foundation of my approach to adversarial techniques in conversational systems. The remaining four principles from Stajano and Wilson's framework (Distraction, Herd, Dishonesty, and Kindness) were deemed less directly applicable to text-based conversational contexts. Distraction, for instance, typically relies on sensory overload or physical misdirection, which is difficult to implement in purely text-based interactions. Similarly, the Herd principle (following others' behaviour) requires visible group dynamics that are not inherently present in one-on-one chatbot interactions. The Dishonesty principle assumes that users knowingly engage in unethical actions, a scenario less applicable to conversational systems. Finally, the Kindness principle, while relevant in certain contexts, lacks the same level of direct applicability in conversational systems compared to the three selected principles.

Having established the rationale for focusing on Need & Greed, Time Pressure, and Social Compliance, it is important to show how these principles were adapted for use in the experimental studies. Figure 3.2 presents a summary of this process, mapping the selected psychological principles onto their concrete operational forms within text-based conversational systems. This representation provides a bridge between the theoretical

discussion and the methodological design, and it sets the stage for the predictions outlined in the next section.

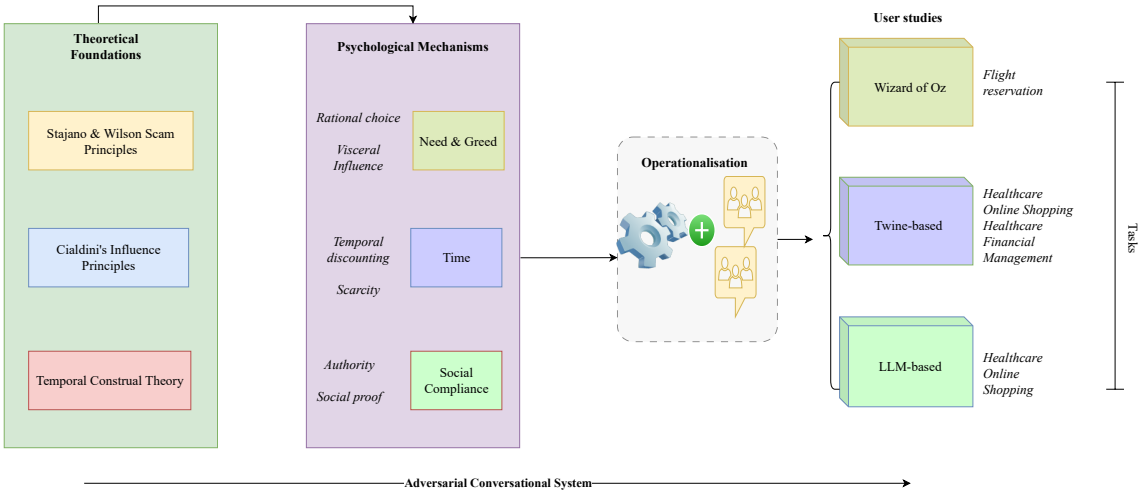


Figure 3.2: Framework illustrating the operationalisation of theoretical principles within adversarial conversational systems. Foundational theories (Stajano & Wilson's Scam Principles, Cialdini's Influence Principles, and Temporal Construal Theory) were translated into three psychological mechanisms—Need & Greed, Time, and Social Compliance—which informed the design of user studies (Wizard of Oz, Twine-based, and LLM-based) across various task domains to examine manipulation and user response patterns.

3.1.2 Expected outcomes and theoretical predictions

Based on the theoretical frameworks discussed in section 2.2, the following predictions detail how user responses to adversarial techniques might vary. These predictions help to link theoretical assumptions directly to the practical outcomes observed in chapter 4, chapter 5 and chapter 6.

General predictions across all techniques

From my integration of Stajano and Wilson's scam principles, Cialdini's influence principles, and Temporal Construal Theory, a set of general predictions can be formulated regarding user responses to adversarial techniques.

Effectiveness of multiple principles When multiple principles are combined, they should produce stronger effects than any single principle in isolation, as the psychological mechanisms may complement each other. While I did not test combinations within individual studies, the progression across studies offers insights into how these techniques compare in effectiveness.

Individual differences Users with higher digital literacy and security awareness should demonstrate greater resistance to all adversarial techniques, reflecting their enhanced ability to recognise manipulation attempts.

Recognition-behaviour gap Users may recognise that they are being manipulated but still comply with requests, particularly when attractive benefits are offered (Need & Greed) or when authoritative sources request information (Social Compliance). This prediction comes from Loewenstein's visceral influence theory, which suggests that immediate psychological influences can override cognitive awareness [48].

Temporal evolution The effectiveness of techniques should change over extended interactions, with users potentially developing resistance strategies as conversations progress and their awareness increases. This prediction was particularly relevant to the LLM-based user study (chapter 6), where more dynamic, extended interactions were possible.

Technique-specific predictions

Each of the adversarial techniques that I used has its own predictions based on the theoretical frameworks.

Need & Greed prediction

- The effectiveness of Need & Greed should vary based on the perceived value-to-risk ratio of the exchange, with higher-value incentives resulting in greater compliance [47].

- Users should be more willing to disclose low-sensitivity information for modest benefits but require substantially greater incentives for high-sensitivity information [145].
- Based on rational choice theory, when perceived benefits outweigh the perceived risks, compliance rates should be high regardless of other factors [146].
- Following Loewenstein's visceral influence theory [48], immediate and concrete benefits should produce more pronounced effects than delayed or abstract benefits.

Time pressure predictions

- Following Temporal Construal Theory [29], Time pressure should shift users toward concrete, immediate concerns and away from abstract principles like privacy protection.
- Users under time pressure should make less consistent decisions and demonstrate poorer recall of information disclosed, reflecting reduced cognitive processing [24, 25].
- Time pressure effects should be stronger for complex decisions that normally require deliberation than for simple, routine decisions [147].
- Based on scarcity principles [28], the perception that time is running out should increase the perceived value of the offered opportunity, potentially increasing compliance.

Social compliance predictions

- Authority-based compliance should be strongest in domains where expertise is highly valued [127] (e.g., healthcare) and weaker in domains where expertise is less critical.

- Social proof elements should be most effective when users are uncertain about appropriate behaviour, consistent with Cialdini's findings on social validation [28].
- The combination of authority claims with social proof should produce stronger effects than either principle alone [148].
- Users should demonstrate higher trust in systems presenting authority markers but may not recognise the influence of these markers on their behaviour [1].

Context-specific predictions

The theoretical frameworks also suggest that contextual factors should significantly moderate the effectiveness of adversarial techniques.

Domain Sensitivity Adversarial techniques may be less effective in highly sensitive domains (e.g., healthcare, finance), where ethical concerns, patient safety, and the necessity of expert judgment result in increased caution among users [53].

Perceived risk In high-risk contexts, users may be more resistant to manipulation as they assess risks more carefully, weighing potential harms against benefits more cautiously [47].

Unexpected Outcomes and Contradictions

The theoretical frameworks also suggest potential contradictions or unexpected outcomes that might emerge. Authority appeals (social compliance) can backfire in some contexts if users perceive them as lacking authenticity or credibility, potentially triggering reactance or reduced trust rather than compliance [8].

The recognition-behaviour gap might manifest differently across techniques, with users potentially more aware of Need & Greed manipulation yet still susceptible due to strong and immediate visceral influences [48].

Different techniques may interact with domain contexts in unexpected ways, with some techniques remaining effective due to inherent psychological vulnerabilities, even in contexts theoretically associated with greater resistance [1, 28].

These potential contradictions highlight the importance of conducting user studies to clarify how these principles operate specifically in conversational system contexts, where the dynamics may differ from traditional face-to-face interactions. My progression from studying Need & Greed and Time techniques in chapter 4 and chapter 5 to examining Social Compliance and Need & Greed in the LLM-based user study (chapter 6) allows for exploring both expected and unexpected patterns in how these techniques influence user attitudes and behaviour.

3.1.3 Psychological Mechanisms of Influence

The adversarial techniques described above are most effectively understood when situated within established psychological frameworks of persuasion and deception. Stajano and Wilson [1] provided one of the earliest systematic accounts of psychological principles exploited in scams. Their framework identified seven "scam principles" that explain how individuals are manipulated, including Need & Greed, Time Pressure, and Social Compliance. While originally developed in the context of email and face-to-face fraud, these principles remain highly relevant to conversational systems, which replicate many of the same psychological triggers in interactive digital settings.

Cialdini's principles of influence [28] also provide a critical foundation. Reciprocity, authority, scarcity, and social proof have all been observed in conversational contexts. For example, reciprocity is evident when chatbots use friendly greetings or small favours to elicit compliance [7] while authority can be signalled through professional persona design. Scarcity and urgency align closely with Time Pressure techniques, and social proof underpins many Social Compliance strategies. Research on conversational agents confirms that these principles can be embedded subtly in dialogue, influencing disclosure and compliance [7, 37, 12].

Temporal Construal Theory (TCT) further explains the effectiveness of Time Pressure in conversational contexts. TCT suggests that people's decision-making shifts

when faced with near-term versus distant outcomes [29]. Liu et al. [24] and Thones and Oberfeld [38] demonstrate that under time constraints, individuals focus narrowly on immediate consequences, reducing their capacity for reflective, systematic reasoning. This aligns with how chatbots create urgency through countdowns or limited-time offers, pushing users toward heuristic rather than deliberative processing.

Finally, the concept of anthropomorphism interacts with these psychological mechanisms. As Larson [14] and Murtarelli et al. [30] argue, human-like design features can amplify the effectiveness of adversarial techniques by strengthening rapport and perceived legitimacy.

In conversational systems, the psychological levers identified by Stajano, Cialdini, and TCT are therefore not abstract theories but operational principles that can be directly embedded in dialogue to influence users.

3.1.4 Implications for User Vulnerabilities

Together, these psychological mechanisms and adversarial techniques highlight that user vulnerabilities in conversational systems extend well beyond technical exploits. They reveal how trust, disclosure, and compliance are shaped in ways that are difficult for users to recognise or resist. For example, Need & Greed leverages visceral influences to override caution [48], Time Pressure disrupts deliberative reasoning [24, 25], and Social Compliance capitalises on obedience to authority and social norms [144, 27]. These are psychological rather than technical vulnerabilities. They are intangible, embedded in the flow of dialogue, and resistant to conventional security measures. They manifest in patterns of user behaviour that have only recently begun to be empirically examined. Users may recognise manipulation yet still comply, an occurrence conceptualised in this thesis as the *recognition–behaviour gap*. This gap demonstrates that awareness alone does not guarantee resistance as affective and cognitive biases can still drive compliance. Studies of conversational agents further suggest that vulnerabilities may vary by domain. For instance, users may be more cautious in healthcare contexts but more susceptible in retail or social scenarios [12, 4].

These findings highlight the need to treat psychological vulnerabilities as an important component of security and trust in human–AI interactions. Unlike technical vulnerabilities, which can be patched, psychological vulnerabilities are rooted in fundamental human tendencies that adversarial techniques exploit. Therefore, the challenge is understanding these vulnerabilities and anticipating how they may evolve as conversational systems become more sophisticated. This provides the rationale for the empirical studies that follow in chapter 4, chapter 5 and chapter 6, which systematically investigate these vulnerabilities across different domains and implementation technologies.

3.2 Part B: Methodological Approach

This section presents the methodological framework that guided the investigation of adversarial techniques in text-based conversational systems. It introduces the overall research design, explains the platforms through which the experiments were conducted, and outlines how participants were recruited and engaged. It also describes the evaluation metrics used to capture both behavioural and attitudinal responses. Together, these elements provided the foundation for systematically testing the theoretical principles outlined in Part A under conditions that balanced control with ecological validity.

3.2.1 Research Design and Rationale

The research adopted a within-subject design in which each participant engaged with multiple scenarios across different experimental conditions. This approach allowed direct comparison of responses to adversarial techniques versus control conditions while reducing the variability associated with between-subject differences. A staged progression across three studies was used, beginning with controlled Wizard of Oz simulations, then interactive narrative scenarios, culminating in dynamic LLM-based interactions. This design was chosen to balance control and ecological validity. The early studies provided tight experimental manipulation, while later studies captured the adaptive and unpredictable features of real-world conversational systems. By systematically increas-

ing the sophistication of the implementation, the design made it possible to examine not only whether adversarial techniques work, but also how their effectiveness shifts across platforms and contexts.

The research progression employed a deliberate strategy in selecting which adversarial principles to investigate at each stage. The first two studies (chapters 4 and 5) both examined Need & Greed and Time techniques, whilst the third study (chapter 6) replaced Time with Social Compliance alongside Need & Greed. This sequencing was purposeful rather than arbitrary.

Need & Greed and Time were selected for both foundational studies to establish baseline effects and enable direct comparison across implementation technologies (WoZ and Twine) before introducing additional technique variation. Both techniques are readily operationalised through relatively standardised manipulations. Need & Greed involves offering explicit incentives, whilst Time imposes temporal constraints, making them well-suited to the scripted and semi-scripted implementations of the early studies. Maintaining consistency in the adversarial techniques across these first two studies allowed the research to isolate the effects of implementation technology and domain context without confounding these factors with simultaneous changes in the psychological mechanisms being tested.

The introduction of Social Compliance in the LLM-based study (chapter 6) served multiple objectives. First, it broadened the range of psychological mechanisms investigated beyond reward-based and temporal manipulations to include authority-based influence. Second, Social Compliance techniques proved particularly well-suited to the dynamic capabilities of LLM-based systems, which can generate authoritative language patterns and contextually appropriate expertise demonstrations in ways that scripted systems cannot. Third, this progression enabled the research to explore how different implementation sophistication levels affect different types of adversarial techniques. Earlier studies established how reward and urgency operate in controlled settings, whilst the LLM study revealed how authority-based manipulation manifests in naturalistic interactions. This staged approach to principle selection thus balanced the need for methodological consistency with the opportunity to investigate diverse psychological

vulnerabilities across the thesis.

Across all three studies, I employed a within-subject experimental design where participants experienced multiple experimental conditions. This established approach in psychology and behavioural sciences helps control for individual differences that might obscure results in between-subject designs. By having participants serve as their own controls, this design reduces the variability of individual differences, increasing statistical power [149, 150].

In my research, this approach offered key advantages. It required fewer participants to achieve statistical power compared to between-subject designs, which is particularly valuable given the complexity of implementing different adversarial techniques. In addition, it allowed for direct observation of how individual participants responded to various techniques within each study.

However, this design required careful management of potential carryover effects, where experiencing one condition might influence responses to subsequent ones. To address this, I followed Miller's [151] recommendations and used a Latin square design to randomise the order of conditions and scenarios across all studies.

3.2.2 Experimental Platforms

Wizard of Oz Framework

The Wizard of Oz (WoZ) framework as shown in Figure 3.3 is utilised in HCI and user experience design. This framework allows a researcher to simulate the functionality of a system that is not yet fully developed by having a human (wizard) control the system's responses in real-time.

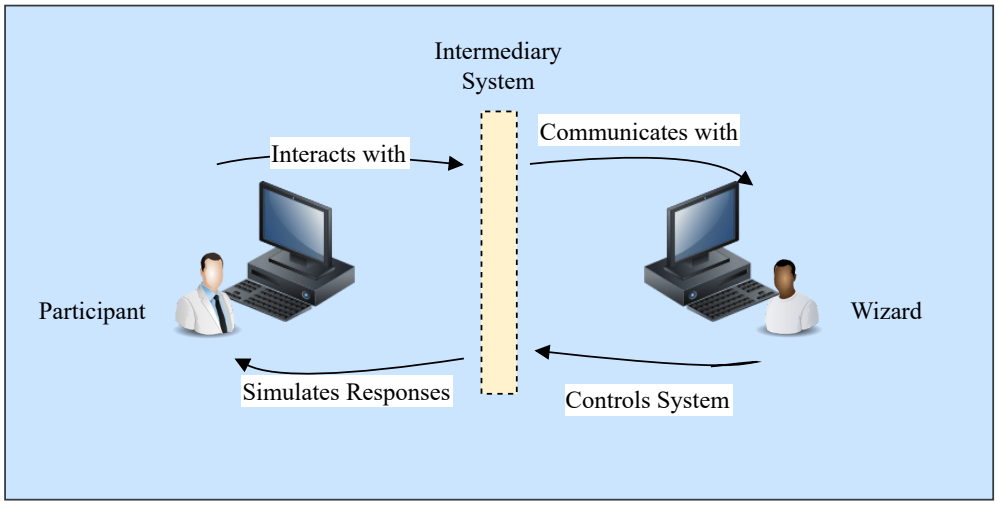


Figure 3.3: Schematic representation of the Wizard of Oz framework used in the first experimental study. The participant interacts with an intermediary system, unaware that a human 'wizard' controls and simulates the system's responses, enabling realistic conversational exchanges while maintaining experimental control over dialogue content and flow.

This framework is valuable for evaluating user interactions with systems like chatbots or any other conversational agents as it helps provide insights into system performance and interaction dynamics without the need for a fully functional prototype. As I discovered in my investigation of adversarial techniques in conversational systems, this methodology provided me invaluable insights into user attitudes.

Historically, the origins of the WoZ methodology can be traced to early HCI research, where it served as a prototyping tool for testing interactive systems before full technological implementation. This approach enables researchers to gather valuable user feedback on system functionality that would otherwise require extensive resources to implement [152, 153]. The framework's flexibility facilitates rapid iteration of design concepts, allowing for detailed exploration of user interactions with conversational agents in controlled environments [154].

The methodology has evolved significantly from its origins in simple response simulation. While early implementations focused primarily on basic dialogue exchanges,

modern applications encompass complex multi-turn interactions, emotional response simulation, and context-aware behaviour modelling [155, 156]. This evolution mirrors the increasing sophistication of conversational systems themselves, making the WoZ framework particularly suitable for investigating how adversarial techniques might influence user behaviour.

In my research, the WoZ methodology proved especially effective for examining how adversarial techniques influence user attitudes in conversational systems. By carefully controlling and observing user responses to different conversational strategies, I could investigate the impact of Need & Greed and Time techniques while maintaining experimental validity. This approach has proven particularly valuable in similar contexts, such as developing health-related chatbots, where understanding patient interactions is crucial for creating effective support systems [157, 158].

Implementing the WoZ study required careful attention to methodological considerations, particularly in experimental design and data collection. As the wizard (I played this role), I developed comprehensive guidelines to maintain standardised responses while preserving natural conversational flow across all experimental conditions. This standardisation proved crucial for ensuring consistency in how adversarial techniques were presented to participants. I established clear communication channels between the wizard and participants to achieve this consistency, implemented predefined response protocols, and incorporated real-time monitoring capabilities. These components worked with my data collection mechanisms, capturing quantitative metrics and qualitative insights through questionnaires administered during and after the experiment using Qualtrics[1]. This balance proved essential in maintaining the illusion of system autonomy while gathering meaningful research data [58, 159].

However, implementing the WoZ methodology presented several significant challenges that required careful consideration and mitigation strategies. The potential for operator bias emerged as a primary concern, as it could affect the authenticity of simulated interactions [160, 161]. Furthermore, maintaining consistent response patterns across multiple sessions required careful attention to protocol adherence. The technical imple-

[1] Qualtrics: https://www.qualtrics.com

mentation required a very good real-time communication system, while ethical considerations raised important questions about informed consent and research transparency.

I addressed these challenges through careful experimental design and standardised response protocols, ensuring consistency across all participant interactions and strict compliance with the ethical guidance provided by the CIS Departmental Ethics Committee.

Twine Platform

Building upon insights from the WoZ study, Twine[2] served as the primary tool to develop interactive scenarios in my second investigation (chapter 5). This platform offers a practical approach to creating branching conversational flows, presenting a visual interface where dialogue segments, called 'passages' (Figure 3.4), can be linked together to form complex interactions [162]. This structure proved particularly useful for implementing adversarial techniques across different scenarios.

The platform offered several key advantages for the implementation of my research. First, its visual development environment allowed for a clear organisation of conversational pathways. Each passage represented a distinct interaction point, making it straightforward to map out how different adversarial techniques would manifest within scenarios. This visual approach, as noted by [162], helps to bring a spatial dimension to the development process, ensuring consistent implementation of Need & Greed and Time techniques across all experimental conditions.

Secondly, Twine's built-in state management capabilities enabled tracking of user choices throughout each scenario [163]. This feature was essential to maintain contextual awareness and ensure participant interactions remained coherent across different stages of the experiment. The system could adapt its responses based on previous user choices, creating more realistic conversational flows.

Third, the platform generates standard HTML output, which is seamlessly integrated with our experimental web-based setup [164]. This compatibility simplified

[2]Twine: https://twinery.org

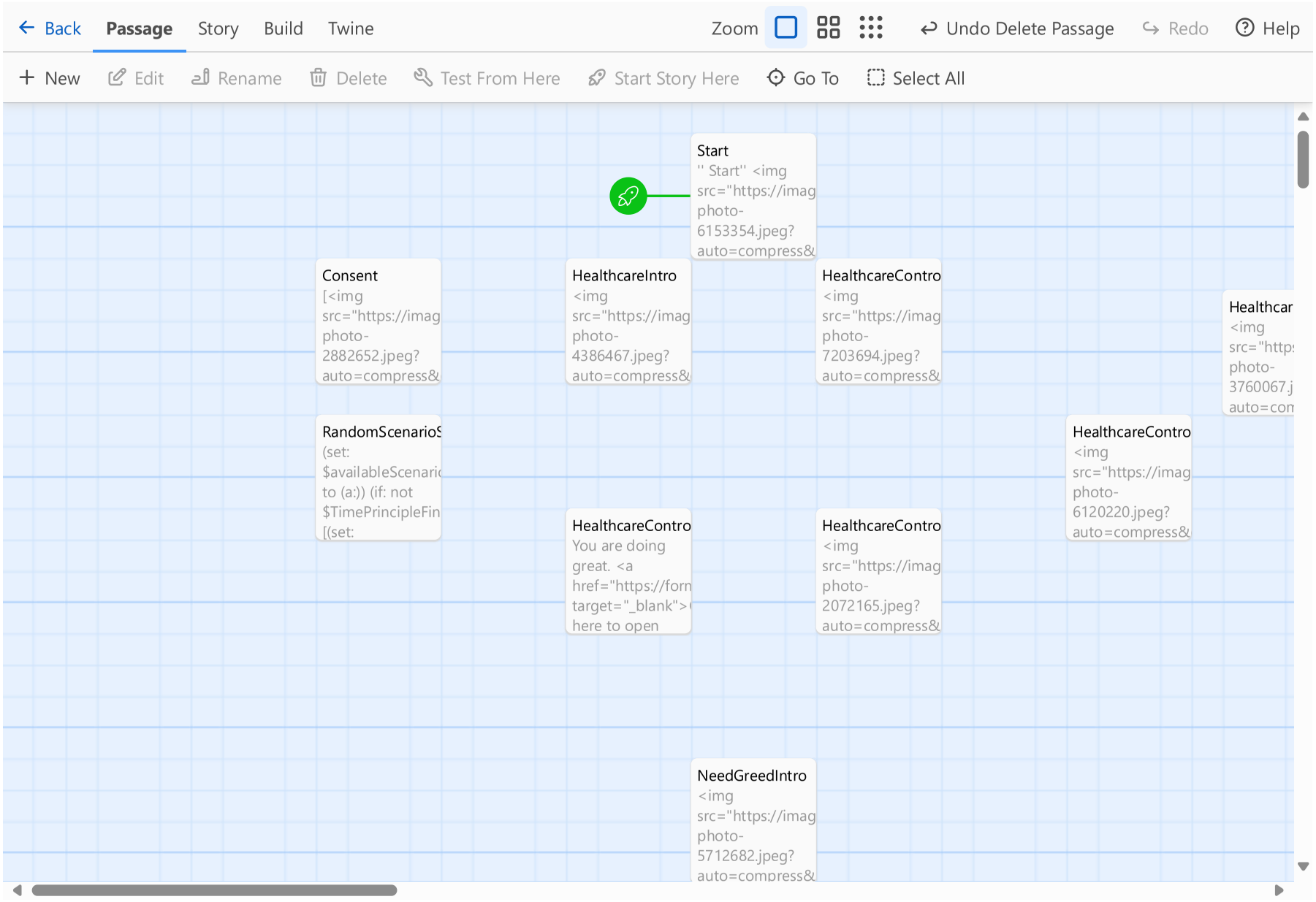


Figure 3.4: Twine Implementation showing the branching narrative structure for experimental scenarios. The visual interface displays the interconnected passages representing the healthcare scenario pathways, including consent procedures, introduction screens, and the three experimental conditions (Control, Need & Greed, and Time Principle). Each node represents a discrete interaction point, with connecting lines illustrating the participant's potential navigation routes through the conversational

the deployment and allowed participants to access scenarios through standard web browsers, reducing technical barriers to participation. Finally, Twine's straightforward implementation approach and support for custom JavaScript and CSS [164] provided the flexibility needed for my experimental design. I could readily incorporate questionnaires between scenarios and implement randomisation for scenario presentation. These capabilities made Twine particularly suitable for my research needs. It allowed me to create consistent, controlled scenarios while maintaining the natural flow of conversation necessary for examining user attitudes and behaviours.

While Twine was valuable for creating structured scenarios, its successful implementation revealed the need for more dynamic interaction capabilities. The platform's predetermined pathways, though effective for controlled experiments, highlighted opportunities for exploring more sophisticated system-user interactions. This recognition led to the development of a more flexible solution using the Flask framework with integrated LLM capabilities.

This progression from Twine to Flask represented a significant advancement in how adversarial techniques could be implemented and studied. While Twine excelled at creating branching narratives across healthcare, online shopping, financial management, and home security scenarios, its pre-scripted nature limited the exploration of more nuanced user responses. The need to understand how users might react to more dynamic and adaptive adversarial techniques motivated the transition to a more sophisticated implementation.

The experience gained from Twine's implementation directly informed the design choices in the Flask application. Understanding how users navigated structured conversational flows helped identify which aspects of interaction needed greater flexibility and responsiveness. This insight proved particularly valuable when integrating Llama 3.3-70b, as it helped balance the need for experimental control with the desire for more natural conversation patterns.

Flask Application

The Flask[3] application marked a significant shift from our earlier methodological approaches. Unlike the structured Twine pathways or the human-operated WoZ framework, this implementation introduced real-time behavioural metrics and scenario-specific adversarial techniques. The three core dependencies of the flask framework: routing, debugging, and the web server gateway interface (WSGI) [165].

For this study, the Flask application was designed to facilitate real-time interactions between participants and the adversarial chatbot, using Llama 3.3-70b to generate contextually relevant responses. The application architecture consists of three primary components: the backend, the frontend interface and the data capture and logging system. This comprehensive design enabled more sophisticated implementation of adversarial techniques while maintaining experimental control.

The backend served as the core processing unit, managing the communication between the frontend interface and the system's processing layers. It integrates Llama 3.3-70b to generate chatbot responses customised to participant input during interactions. As shown in Listing 3.1, the `/message` route handled `HTTP` requests, dynamically processing user queries and generating responses in real time. This ensured a seamless and efficient interaction between the user and the system.

Listing 3.1: Flask route handling user-bot interactions and integrating real-time behavioural metrics.

```
@app.route('/message', methods=['POST'])
@limiter.limit("5 per minute")
def add_message():
    """Handle message exchanges between user and bot"""
    try:
        data = request.json
        if not data or 'session_id' not in data or 'content' not
            in data:
```

[3]Flask: https://flask.palletsprojects.com

```
            return jsonify({"error": "session_id and content are
                required"}), 400

        session_id = data['session_id']
        content = data['content']
        current_session = db.session.get(Session, session_id)
        if not current_session:
            return jsonify({"error": "Invalid session ID"}), 400

        start_time = time.time() * 1000
        user_message = Message(
            session_id=session_id,
            content=content,
            is_bot=False,
            timestamp=start_time
        )
        db.session.add(user_message)

        bot_response = get_bot_response(session_id, content,
            current_session.task.principle)
        end_time = time.time() * 1000
        response_latency = end_time - start_time

        bot_message = Message(
            session_id=session_id,
            content=bot_response,
            is_bot=True,
            response_latency=response_latency,
            timestamp=end_time
        )
        db.session.add(bot_message)

        metric = db.session.execute(select(Metric).filter_by(
```

```
            session_id=session_id)).scalar_one_or_none()
        if not metric:
            metric = Metric(
                session_id=session_id,
                scenario=current_session.scenario,
                condition=current_session.condition,
            )
            db.session.add(metric)

        metric.total_messages += 2
        metric.response_latency = (metric.response_latency +
            response_latency) / metric.total_messages
        db.session.commit()

        return jsonify({"bot_response": bot_response}), 200
    except Exception as e:
        logger.error(f"Error in add_message: {str(e)}", exc_info=
            True)
        return jsonify({"error": "An internal error occurred.
            Please try again."}), 500
```

The frontend presented users with an easy-to-use interface featuring a text input field for queries and a response display area for system-generated replies. This design prioritised simplicity and usability, ensuring participants could focus on the interaction without distractions. Real-time communication between the front- and back-end was achieved through asynchronous `HTTP` requests, enabling instant responses and dynamic updates that enhanced the natural flow of conversation.

An essential feature of the application was its ability to log user interactions automatically. Behavioural metrics such as response times, disclosure and compliance rates, and user engagement were captured at the backend automatically. This logging functionality helped in the collection of user interaction data for analysis, complementing the self-reported attitudes collected through post-task questionnaires.

The application was deployed on Heroku, a cloud-based platform that provided a stable and reliable environment for hosting and real-time processing. Heroku's scalable architecture ensured smooth performance even under varying levels of user engagement. Using this deployment strategy made the application easy for everyone to access on any device, making sure it worked well and looked the same for all participants.

This implementation offered several key advantages over previous methodological approaches. Unlike the WoZ framework's human-operated responses or Twine's pre-scripted pathways, the Flask application enabled dynamic conversations through the LLM I integrated. This capability allowed for examining how users respond to adversarial techniques in more naturalistic interactions while still maintaining necessary experimental controls. Another significant advancement was the real-time data collection. While earlier approaches relied heavily on post-interaction questionnaires, this system captured behavioural metrics during the actual interaction. Response latencies, interaction patterns, and information disclosure rates could be measured precisely, providing a deeper understanding of how users engage with adversarial conversational systems.

This implementation continued to benefit from our established participant recruitment through Prolific Academic, which had proven effective during the Twine-based study (chapter 5). The increased complexity of interactions and data collection in the Flask implementation reinforced the value of Prolific's pre-screening capabilities and diverse participant pool. Building on my experience with participant recruitment from the Twine study, I maintained detailed selection criteria while adapting experimental protocols to accommodate the more dynamic nature of the Flask-based interactions.

3.2.3 Participants and Recruitment

Prolific Academic[4] (ProA) is an online platform used for the recruitment of participants for academic research. Having successfully used Prolific Academic in both twine- and flask-based studies, this platform proved instrumental in maintaining consistent participant recruitment standards. Compared to other recruitment tools like Amazon

[4]Prolific: https://www.prolific.com

Mechanical Turk[5], Prolific stands out for its emphasis on participant diversity, fairness, and data reliability [166, 167].

One key advantage of Prolific is its robust pre-screening capabilities, which enable careful participant filtering based on demographic criteria such as age, gender, educational background, and country of residence. This feature proved particularly valuable across both the Twine and Flask studies, ensuring consistent participant selection while maintaining sample diversity. Participants recruited through the Prolific platform consistently produced higher-quality responses and demonstrated lower attrition rates compared to other platforms like Amazon MTurk [166].

However, it's important to acknowledge certain limitations in using Prolific. Although the platform supports participant diversity, its user base remains primarily concentrated in Western countries, which can limit the generalisability of the findings to the global population [168]. A recent platform update has partially addressed this through the "Bring your own participant" tool. However, this feature was not used in my Twine-based study as it became available after the data collection phase.

The consistency in participant recruitment through Prolific complemented my use of within-subject design across all the user studies, from the initial WoZ framework through to the Flask implementation. This consistency in method was important for maintaining experimental validity while also examining the increasingly complex aspects of the adversarial techniques I used.

3.2.4 Evaluation Metrics

I used a structured evaluation framework to assess the psychological vulnerabilities created by adversarial techniques in the three studies (chapter 4, chapter 5 and chapter 6). This approach was essential for characterising how these techniques affect users across different contexts and implementation technologies, a central aspect of my thesis statement. The framework enabled consistent measurement of user attitudes and behaviours while supporting methodological progression as my research evolved from controlled to increasingly naturalistic implementations.

[5] Amazon Mechanical Turk: https://www.mturk.com

Attitudinal Metrics

Four attitudinal metrics were consistently measured in the three studies.

Trust This measured how confident users felt about the system's reliability, integrity and intentions. A model of trust in conversational agents examines chatbot-related, environment-related, and user-related trust factors [8]. This metric was operationalised through 5-point Likert scale questions such as "How much did you trust the conversational agent's advice or suggestions in this interaction?" (1: Not at all - 5: Completely). While Nordheim et al. used a 7-point scale with items like "I experience this chatbot as trustworthy," I adapted this to a 5-point scale for consistency with other measures. This metric was crucial for understanding how adversarial techniques might affect trust, which is a fundamental aspect of human-computer interaction, and how its effects could differ between contexts, especially since trust plays such an important role in whether people adopt and continue using conversational technologies [169].

Comfort Level This metric assessed users' perceived ease and psychological comfort while engaging with the conversational system. This was measured through questions such as "How comfortable did you feel during your interaction with the conversational agent?" (1: Very uncomfortable - 5: Very comfortable). Users reported significantly less frustration and higher satisfaction with conversational interfaces compared to traditional web interfaces, particularly amongst those with lower digital literacy [86]. This aligns with findings that emotional and relational benefits of disclosure were equivalent whether participants thought they were disclosing to a chatbot or a person [81], supporting the Computers as Social Actors (CASA) framework. Comfort represents an important dimension of user experience that adversarial techniques may significantly affect, as manipulations that undermine perceived understanding, positive relational dynamics, or system usability could negatively impact psychological ease during interaction, potentially leading to changes in what users choose to share.

Willingness to disclose information This metric evaluated the willingness of users to share personal or sensitive information with the system. Research on self-disclosure in conversational agents found that emotional and relational benefits of disclosure were equivalent whether interacting with perceived chatbots or humans [81]. This was assessed through context-specific questions such as "The offered benefits made me willing to share information" (1: Strongly disagree - 5: Strongly agree). The communication context can significantly influence the willingness to disclose [11], with functional contexts (such as professional health advice) encouraging more willingness to share than social contexts. I adapted questions about specific types of personal information to measure participants' willingness to disclose. This metric directly measured the behavioural impact of adversarial techniques on information security, providing insight into how psychological manipulation translates into security risks. This is particularly important because adversarial techniques can alter how users perceive the conversation context or reduce their comfort, potentially altering sharing patterns in ways that put information security at risk.

Perception of privacy and security This metric examined users' beliefs about the safety and confidentiality of their interaction. The privacy calculus model suggests users weigh privacy concerns and risks against potential benefits when making disclosure decisions [170]. I operationalised this through questions such as "I felt concerned about my privacy during this interaction" and "I was worried about the security of my information" (1: Not at all - 5: Very much). These questions directly assessed perceived risk, which Dinev and Hart identified as a significant barrier to information disclosure. This metric helped me distinguish between what users actually did and what they thought about risk, showing how different contexts might affect their perception. This was essential because adversarial techniques could affect trust (which Dinev and Hart [170] highlight as important for disclosure decisions) without immediately changing users' behaviour, potentially creating security vulnerabilities that would not be obvious just by observing what users did.

Behavioural Metrics

As the research progressed to more sophisticated implementations, additional behavioural metrics were incorporated to capture actual user responses rather than self-reported attitudes in chapter 6.

Compliance rates Compliance was measured as the percentage of successful information elicitation attempts across different categories of requests. An examination of compliance across request categories such as authority-based and social proof [7] informed my analysis here. I further categorised compliance by sensitivity level, operationalising sensitivity by assigning a higher risk value (1.2) to healthcare scenarios due to the sensitive nature of health-related disclosures, compared to online shopping scenarios, which were assigned a lower risk value (1.0). This categorisation facilitated analysis of how compliance varied according to both adversarial techniques and scenario context.

Information disclosure patterns User disclosures have been classified into categories such as general personal disclosures, demographic details, personal preferences, and topic-specific disclosures (e.g., healthcare or shopping-related) [81]. I adopted this framework, quantifying disclosure patterns systematically using NLP-based methods and regular expression matching. Voluntary disclosures beyond direct conversational prompts were also measured to capture spontaneous user openness under adversarial conditions. Importantly, the categorisation of disclosures was carefully designed not to capture or retain personally identifiable information (PII), in line with ethical standards.

Resistance strategies. A privacy-protective framework for identifying resistance tactics [145] provides the foundation for tracking explicit information withholding, rejection of information requests, deflection of system prompts, and the use of security or privacy-related terms. I used NLP pattern detection to identify such strategies in conversational logs.

Recognition indicators Recognition of manipulation attempts was indirectly inferred [112], measured through the increased use of privacy and security terminology, explicit information withholding, and privacy concerns indicators. These behavioural indicators provided insight into implicit cognitive recognition and highlighted the recognition-behaviour gap, where awareness of manipulation may not translate into active resistance behaviours.

User Engagement User engagement was operationalised explicitly as the intensity and extent of interactions, measured through the total number of conversational exchanges (utterances) between participants and the conversational agent. This was quantified as an `engagement_score`, reflecting participants' conversational involvement and providing important behavioural context about user receptivity or resistance to adversarial manipulations, aligning with the conceptualisation of user engagement [171].

Qualitative Evaluation Metrics

Qualitative evaluation metrics were integrated alongside quantitative data to enhance psychological insights into user experiences and behaviours.

Refined Thematic Analysis To complement the quantitative measures, I performed a reflexive thematic analysis of the responses to the post-task questionnaire from chapter 5 following a structured methodology [172]. This involved systematically coding user responses, followed by theme generation, review, and refinement. The resulting themes provided deeper psychological insight into the vulnerability patterns uncovered by quantitative measures, revealing the underlying cognitive and emotional mechanisms.

Contextual Interpretation Post-task questionnaire responses were contextually analysed in healthcare, financial and online shopping scenarios, recognising that similar behaviours can have different implications depending on the domain's sensitivity. This approach helped us better understand how contexts in adversarial techniques influenced

user vulnerabilities and behaviour.

Temporal and Contextual Integration Qualitative insights were integrated with behavioural data to explore how different adversarial strategies affected users over time and across contexts. This combination clarified how adversarial techniques functioned differently across interaction timelines and scenarios, specifically in chapter 6, providing further insights into user behaviour when exposed to these techniques.

3.3 Summary

This chapter has provided both the theoretical and methodological foundations of the thesis. Part A introduced the psychological principles and adversarial techniques underlying the study of conversational system vulnerabilities. Part B described the methodological design and experimental platforms used to systematically investigate these principles. This chapter lays the foundation for empirical studies in subsequent chapters by linking theoretical insights to methodological implementation. The next chapter applies these foundations to the first user study, which examines how adversarial techniques manifest in a controlled Wizard-of-Oz environment.

Chapter 4

A Foundational Study of Adversarial Techniques in Conversational Systems: Impacts on User Attitudes

Do not be anxious about anything, but in every situation, by prayer and petition, with thanksgiving, present your requests to God. Phil 4:6 (NIV)

Building on the background established in chapter 2 and the theoretical foundations outlined in chapter 3, this chapter presents a foundational investigation into how adversarial techniques influence user attitudes in conversational systems. I focus on two specific adversarial techniques-Need & Greed and Time, examining their effects in a controlled experimental setting.

4.1 Introduction

This chapter provides foundational insights into user attitude patterns and the effectiveness of Need & Greed and Time adversarial techniques through human-operated system simulation. I employed the Wizard of Oz (WoZ) framework to simulate realistic interactions while investigating how these techniques influence key user attitudes: trust, comfort level, willingness to disclose information, and perceived privacy and security. The WoZ approach is a well-established method in HCI research for simulating system

behaviour before full automation [173, 174, 175, 176]. It is particularly effective for natural language interfaces because it enables the study of realistic user interactions while maintaining experimental control. Riek [160] emphasised its suitability when behaviours are feasible for human operators but not yet robustly automated, precisely for controlled adversarial manipulations. In this study, WoZ ensured consistent presentation of Need & Greed and Time techniques while participants experienced fluid dialogue, striking a balance between control and ecological validity. This made WoZ the most suitable method for exploring adversarial techniques at this foundational stage.

4.2 Study Aim, Objectives, and Research Question

This section presents the study-level aim, objectives, and research question for the foundational Wizard of Oz study reported in this chapter. The study-level objectives below address dissertation-level Objectives 1–2 and Research Question 1, as set out in Section 1.4.

4.2.1 Aim.

To determine how Need & Greed and Time adversarial techniques influence user attitudes in text-based conversational systems.

4.2.2 Objectives.

1. Implement Need & Greed and Time techniques in a Wizard of Oz conversational framework.
2. Measure their effects on trust, comfort, willingness to disclose information, and perceptions of privacy and security.
3. Compare responses across conditions within participants to isolate the specific impact of each technique.
4. Establish attitudinal baselines for later studies with broader domains and technologies.

4.2.3 Research Question.

RQ 4.1: How do Need & Greed and Time adversarial techniques influence user attitudes—specifically trust, comfort, willingness to disclose information, and perceptions of privacy and security—in text-based conversational systems?

4.2.4 Contributions.

- Provides empirical evidence that scam-derived adversarial techniques alter key user attitudes in dialogue.
- Validates a procedure for operationalising adversarial cues in conversation while maintaining ecological validity.
- Establishes attitudinal benchmarks that guide hypotheses and design choices for subsequent studies.

The rest of this chapter is organised as follows. section 4.3 outlines the methodology comprising the experimental design, participant recruitment process, and task descriptions. It also details the independent and dependent variables investigated and the data collection methods and control measures employed to ensure the study's validity and reliability. Following this, section 4.4 presents the findings of the investigation, focusing on the impact of Need & Greed and Time adversarial techniques on key user attitudes, with analyses supported by statistical tests and visualisations. The chapter concludes with section 4.5, which synthesises the findings, discusses their implications for the ethical design of conversational systems, and identifies directions for future research.

4.3 Methodology

4.3.1 Experimental Design

I employed a within-subject experimental design to investigate the impact of adversarial techniques on user attitudes in conversational systems. This approach was chosen for

its ability to control for individual differences and increase statistical power with a smaller sample size. Each participant experienced all three experimental scenarios: Control, Need & Greed, and Time, allowing for a direct comparison of responses across conditions.

The within-subject design offers several advantages in this context. Firstly, it reduces the impact of individual variability, as each participant serves as their own control. This is particularly important when studying user attitudes, which can vary significantly between individuals. Secondly, it allows for a more efficient use of participants, enabling me to gather more data with fewer subjects. Lastly, this design enhances my ability to detect subtle differences in user responses across scenarios, which is crucial given the potentially nuanced effects of adversarial techniques.

4.3.2 Participants

Sample Size Determination

I conducted an a priori power analysis using G*Power 3.1 to determine the appropriate sample size. Based on previous studies in conversational agent interactions [177], I anticipated a medium effect size ($f = 0.25$). The analysis indicated that a minimum of 23 participants would be required to achieve 80% power with an α level of 0.05 for repeated measures ANOVA with three measurements.

Recruitment

Participants were recruited through convenience sampling using social media platforms (WhatsApp) and word-of-mouth referrals within the University of Strathclyde community. Recruitment materials included a brief study description outlining the general nature of the research (investigating user interactions with conversational systems) without revealing the specific focus on adversarial techniques to avoid priming effects. Interested individuals were provided with detailed participant information sheets and screened for eligibility based on the following criteria: (1) age 18 years or older, (2) proficiency in English, and (3) prior experience using online booking systems or con-

versational agents. No exclusion criteria were applied beyond these basic requirements.

Sample Characteristics

The final sample consisted of 24 participants (12 female, 12 male) aged 18 to 54 years ($M = 31.72$, $SD = 10.11$). The sample represented a highly educated demographic, with the majority (45.8%) holding advanced degrees. Detailed demographic characteristics are presented in Table 4.1.

Table 4.1: Demographic characteristics of participants in the Wizard of Oz study (N=24). The table presents the distribution of gender (50% male, 50% female), age ranges (18–54 years, M=31.72, SD=10.11), and education levels. The sample was highly educated, with 45.8% holding Master's degrees and 33.3% holding Bachelor's degrees, reflecting the convenience sampling approach through the University of Strathclyde community.

Characteristic	**Category**	**N**	**%**
Gender	Male	12	50%
	Female	12	50%
Age	18–24	4	16.7%
	25–34	14	58.3%
	35–44	3	12.5%
	45–54	3	12.5%
Education	Associate degree	2	8.3%
	Bachelor's degree	8	33.3%
	Master's degree	11	45.8%
	Some college, no degree	3	12.5%
Total		24	100%

Pilot Testing

Two pilot rounds were conducted before the main study to refine the experimental protocol and identify potential technical or procedural issues. Each pilot round involved five participants who were not included in the final sample.

Pilot 1: The initial pilot revealed two significant issues. First, participants reported slow response times from the wizard, which disrupted the natural conversational flow and raised suspicions about the system's authenticity. Second, a technical issue with chat link generation resulted in broken links for some participants. These problems

were addressed by refining the wizard response protocol to reduce latency and fixing the technical implementation of the Y99 platform link generation system.

Pilot 2: A second pilot round with five different participants tested the revised protocol. All participants completed the procedure successfully without technical difficulties. Response timing was perceived as natural, conversational flow felt authentic, and all data collection instruments functioned correctly. This confirmed that the experimental setup, timing, conversational patterns, and data capture procedures were ready for the main study. No further modifications were necessary following Pilot 2.

4.3.3 Experimental Setup and Apparatus

Platform and Implementation

I utilised a web-based experimental setup to investigate how adversarial techniques influence user attitudes in conversational systems. The study was conducted using the Y99[1] chat platform combined with the Wizard of Oz (WOZ) framework. In this approach, I covertly assumed the role of the conversational agent (the "wizard"), providing responses that appeared automated to participants while maintaining strict experimental control over interaction patterns and adversarial manipulations. This method preserved ecological validity by simulating a sophisticated conversational system while ensuring consistent application of experimental conditions across all participants.

Each participant joined a private chat session on the Y99 platform through individually generated chat links, maintaining privacy and isolating interactions (see Figure 4.1). As the wizard, I used a pre-prepared library of responses tailored to each experimental condition (Control, Need & Greed, and Time), ensuring consistency in language, tone, and interaction patterns. The standardised response library was developed and refined through the two pilot rounds to ensure natural conversational flow while precisely implementing the intended adversarial techniques.

[1] Y99: https://y99.in

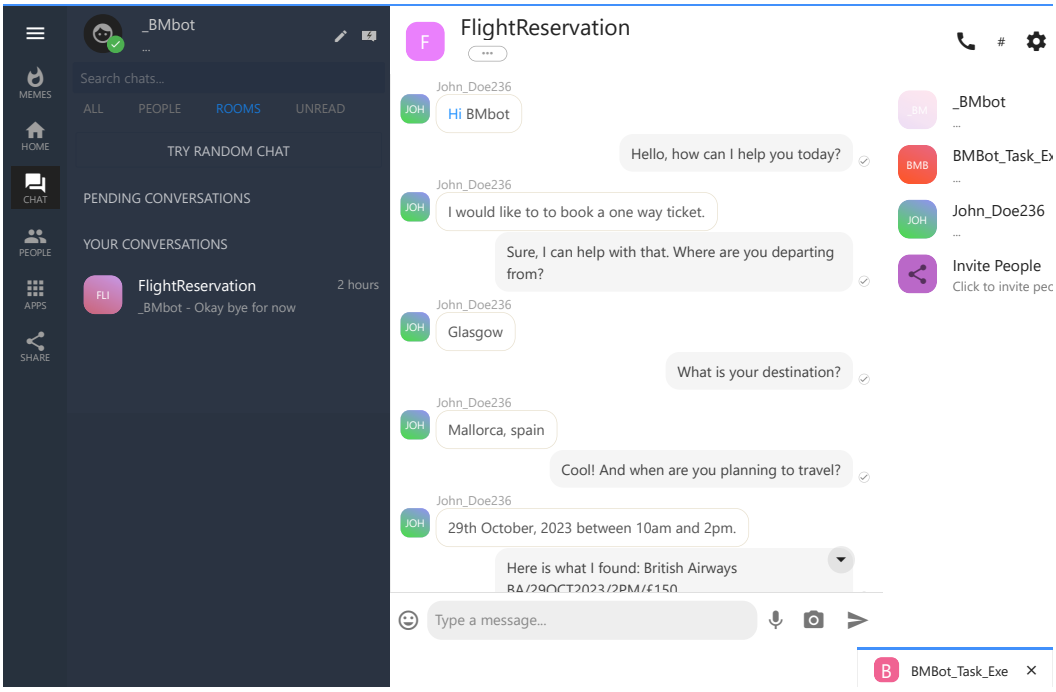


Figure 4.1: Flight reservation task chat interface showing the Y99 platform used for participant interactions.

Rationale for Web-Based Approach

Conducting the study online offered several advantages. The Y99 platform provided a familiar interface similar to real-world messaging applications (e.g., WhatsApp[2], Facebook Messenger[3]), enhancing ecological validity and promoting natural user behaviour. The remote nature of the study allowed participants to engage from comfortable, familiar environments at their convenience, reducing potential situational influences while maintaining research rigour. Additionally, this approach facilitated efficient data collection and reduced logistical challenges associated with laboratory-based studies.

[2]WhatsApp: https://www.whatsapp.com
[3]Facebook Messenger: https://www.messenger.com

Control Measures and Data Protection

To ensure the validity and reliability of findings, several control measures were implemented:

- **Scenario randomisation:** Using a Latin Square design, the presentation order of scenarios was systematically varied across participants to counterbalance potential order effects and practice effects.
- **Standardised interactions:** The pre-prepared response library ensured that all participants within each condition received equivalent conversational patterns, removing potential experimenter bias from spontaneous responses.
- **Anonymised interaction:** Participants were provided with pre-generated dummy personal information (fictitious names, email addresses, and phone numbers) to use during the flight booking task. This approach protected participant privacy while standardising the perceived risk associated with sharing sensitive information across all experimental conditions.
- **Isolated sessions:** Individual chat links prevented cross-contamination between participants and ensured that each session remained independent.

4.3.4 Task and Experimental Conditions

Task Description

Participants were tasked with planning a holiday flight to Mallorca, Spain, using the conversational agent. The task was designed to simulate a realistic online booking scenario with the following standardised parameters:

- **Route:** Glasgow to Mallorca, Spain
- **Date of departure:** 29th October
- **Time:** Between 10:00 and 14:00

- **Class:** Economy
- **Budget:** £150

These parameters were selected through a systematic process balancing three key criteria: relevance, complexity, and engagement. The **route** (Glasgow to Mallorca) was chosen because Mallorca represents a popular holiday destination for UK travellers, ensuring the scenario felt realistic and personally relevant to participants, whilst Glasgow served as a familiar departure point for the Scottish-based sample. The **date** (29th October) and **time window** (10:00–14:00) were selected to introduce moderate scheduling constraints that required participants to evaluate options actively, rather than simply accepting the first available flight, thus increasing task complexity without becoming unrealistic. **Economy class** was chosen as the most commonly booked cabin class, enhancing ecological validity and ensuring participants could draw on prior booking experiences. Finally, the **budget of £150** was deliberately set at a level that was achievable for a short-haul European flight but tight enough to create meaningful trade-offs between cost, timing, and convenience, thereby encouraging sustained engagement with the conversational agent throughout the decision-making process. Together, these parameters created a constrained yet realistic booking scenario that required meaningful interaction with the conversational agent whilst maintaining experimental control through standardisation. Participants were encouraged to ask questions, seek clarifications, and interact naturally, mirroring typical user behaviour in online booking scenarios.

Experimental Conditions

To investigate how adversarial techniques influence user attitudes, I designed three experimental conditions based on established principles from social engineering and persuasion research [1, 178]:

1. **Control condition:** This baseline scenario involved neutral interactions without any adversarial elements. The conversational agent provided helpful, factual

responses focused on completing the flight booking task efficiently. The agent adhered strictly to ethical practices and made no attempts to manipulate user decisions or extract unnecessary information. This condition served as a reference point against which to evaluate the effects of the adversarial techniques.

2. **Need & Greed condition:** In this scenario, the conversational agent employed incentive-based persuasion techniques to influence participants. The agent offered various benefits such as percentage discounts on flights, exclusive membership benefits, priority boarding privileges, and special travel insurance offers contingent upon sharing additional personal information (e.g., email addresses, phone numbers, date of birth). This approach mimicked real-world applications where systems exploit user desires and needs to elicit personal information beyond what is necessary for task completion [1, 7].

3. **Time condition:** In this scenario, the agent created a sense of urgency by emphasising time-sensitive constraints. Participants were informed that flight availability was rapidly diminishing, special offers would expire within minutes, or that prices were increasing in real-time. Phrases such as "only 2 seats left at this price" or "this offer expires in 5 minutes" were strategically incorporated. This simulated high-pressure decision-making environments, drawing on the psychological principle of temporal discounting where individuals prioritise immediate rewards over long-term considerations [178].

The scenarios were randomised for each participant using a Latin Square design to mitigate order effects and ensure balanced exposure across conditions. Each participant experienced all three conditions, with the order counterbalanced across the sample.

4.3.5 Variables

Independent Variables

The experimental manipulations varied across conditions. For the Need & Greed scenario, the independent variables included:

1. **Perceived benefits from sharing information:** The extent to which participants believed that disclosing personal information would yield tangible advantages (e.g., discounts, improved service).

2. **Perceived risk from sharing information:** Participants' assessment of potential negative consequences (e.g., privacy breaches, data misuse) associated with information disclosure.

3. **Motivation to share information:** The strength of participants' intrinsic and extrinsic drivers to disclose personal data in exchange for offered benefits.

4. **Perceived trust in the system:** Participants' confidence in the conversational agent's data handling practices and ethical behaviour.

For the Time scenario, the independent variables were:

1. **Urgency level of the request:** The degree of time pressure communicated by the conversational agent.

2. **User's perceived available time:** Participants' subjective assessment of how much time they had to make decisions.

3. **Perceived time constraints:** The extent to which participants felt they were operating under restrictive deadlines.

4. **Pressure due to time constraints:** Psychological pressure experienced by participants as a result of the temporal manipulations.

Dependent Variables

The dependent variables were selected to comprehensively measure the effects of adversarial techniques on user attitudes:

1. **Trust in the system:** Participants' confidence in the conversational agent's reliability, data protection practices, and ethical behaviour. Measured using 5-point Likert scales assessing various dimensions of trust.

2. **Comfort level during interaction:** The degree to which participants felt at ease while engaging with the conversational agent. Measured through self-reported ratings of comfort and anxiety during the interaction.

3. **Willingness to disclose information:** Participants' readiness to share sensitive personal information with the conversational agent. Assessed through both self-reported intentions and behavioural observations during the task.

4. **Perceived privacy and security:** Participants' beliefs about the safety, confidentiality, and appropriate handling of information shared during interactions. Measured using established privacy concern scales adapted to the conversational agent context.

Measurement Approach and Statistical Validity

All attitudinal dependent variables (trust, comfort level, willingness to disclose, perceived privacy and security) were measured using 5-point Likert scales administered in post-task questionnaires after each experimental condition. The within-subject repeated measures design, combined with the sample size (n=24) determined through a priori power analysis (G*Power 3.1, medium effect size $f = 0.25, power = 0.80, \alpha = 0.05$), ensured adequate statistical power for detecting medium-sized effects using repeated measures ANOVA. This design inherently controls for individual differences by having each participant serve as their own baseline [151], and the repeated-measures ANOVA automatically partitions between-subject variance, effectively removing the influence of stable individual characteristics from the error term when testing for condition effects [150]. The Latin Square randomisation of condition order further ensured that any order effects were distributed evenly across conditions and could not systematically bias the results. Complete questionnaires showing all measurement items and response scales are provided in Appendices B.1.2, B.1.3, B.1.4 and B.1.5 (chapter 4) and Appendices B.2.2, B.2.3, B.2.4 and B.2.5 (chapter 5).

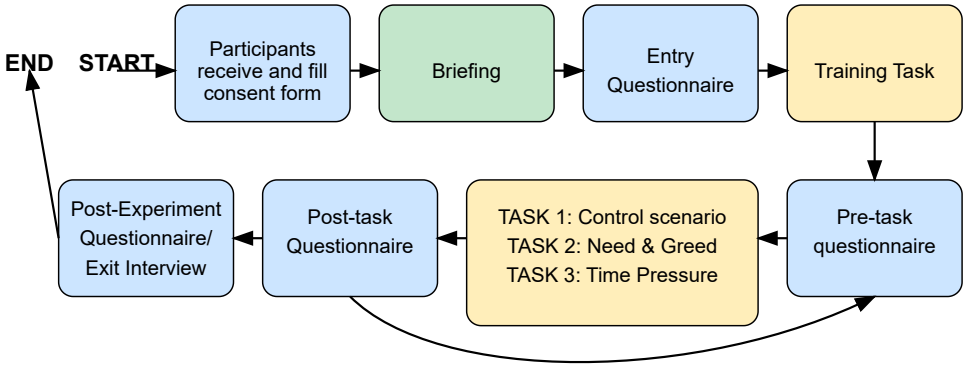


Figure 4.2: Experimental workflow for the Wizard of Oz study showing the complete sequence of participant activities. The flowchart illustrates the progression from initial consent and briefing through entry questionnaire, training task, three experimental interactions (Task 1: Control, Task 2: Need & Greed, Task 3: Time Pressure), post-task questionnaires after each interaction, and concluding with the exit questionnaire. The Latin Square design ensured counterbalancing of condition order across participants to control for carryover effects.

4.3.6 Procedure

The study followed a standardised protocol to ensure consistency across all participants. Participants progressed through six main stages: pre-study information and consent, entry questionnaire, three experimental interactions (with post-task questionnaires after each), and a final exit questionnaire with debriefing. The average study duration was 30–40 minutes. The complete experimental workflow is illustrated in Figure 4.2.

Pre-Study Phase

Upon expressing interest in the study and consent (see Appendix B.1.1), potential participants received a detailed information sheet explaining the study's general purpose (investigating user interactions with conversational booking systems), their rights as participants, data handling procedures, and contact information for the research team and ethics committee. After reviewing this information, participants provided informed consent through an online consent form before proceeding to the study.

Entry Questionnaire

Participants first completed an entry questionnaire designed to gather baseline data. This included:

- **Demographics:** Age range, gender, education level, and English proficiency.
- **Baseline privacy attitudes:** General attitudes toward online data sharing, prior experiences with conversational agents, and typical privacy behaviours when using online booking systems.
- **Technology experience:** Self-reported familiarity with conversational agents, chatbots, and online booking platforms.

These baseline measures characterised the sample's pre-existing privacy attitudes and technology experience, providing important context for interpreting the findings. Although these measures were not formally included as covariates in the statistical analyses, the within-subject experimental design inherently controlled for individual differences by having each participant serve as their own control across conditions. This ensured that the observed effects could be attributed to the experimental conditions rather than to between-subject differences in pre-existing attitudes.

Experimental Interactions

Each participant completed three flight booking tasks, one for each experimental condition (Control, Need & Greed, and Time). The order of conditions was randomised using a Latin Square design to counterbalance potential order and fatigue effects. For each interaction:

1. Participants received task instructions specifying the flight booking parameters (route, date, time, class, budget).
2. Participants were provided with dummy personal information (fictitious name, email address, phone number- see Appendix A.2) to use during the booking process, ensuring anonymity and standardising perceived disclosure risk.

3. Participants accessed their unique Y99 chat link and began interacting with the conversational agent (wizard).

4. The interaction continued until the flight booking task was completed or participants indicated they wished to discontinue.

5. All chat interactions were logged automatically for subsequent analysis.

Post-Task Questionnaires

Immediately after each interaction, participants completed a condition-specific post-task questionnaire. These questionnaires assessed the four dependent variables (trust, comfort, willingness to disclose, perceived privacy and security) using 5-point Likert scales. Questions were tailored to each condition to capture scenario-specific effects. For example:

- In the Need & Greed condition, questions assessed reactions to offered incentives and participants' evaluation of the trade-off between benefits and privacy.
- In the Time condition, questions explored how urgency influenced decision-making and information disclosure.
- In the Control condition, questions focused on baseline perceptions without manipulation-specific elements.

All post-task questionnaires also included open-ended questions allowing participants to describe their experiences, concerns, or notable aspects of the interaction in their own words. The complete post-task questionnaires for the Control, Need & Greed, and Time conditions are provided in Appendices B.1.2, B.1.3 and B.1.4 respectively.

Exit Questionnaire

After completing all three scenarios, participants completed an exit questionnaire that captured:

- Comparative evaluations across the three conditions (e.g., which scenario felt most trustworthy, which created the most pressure).
- Overall reflections on the study experience.
- Suspicion checks to assess whether participants believed they were interacting with an autonomous system or suspected human involvement.
- General feedback and suggestions for improvement.

The exit questionnaire provided valuable context for interpreting quantitative results and revealed insights that structured items might not capture. The complete exit questionnaire is provided in Appendix B.1.5.

Debriefing and Compensation

Following the exit questionnaire, all participants received a comprehensive debriefing that explained the true nature of the study, including the use of the Wizard of Oz methodology and the specific adversarial techniques employed in each condition. This debriefing ensured transparency and allowed participants to ask questions about the research. Participants then received their £10 shopping voucher as compensation. The debriefing process also served as an opportunity to address participants' concerns about their data or the experiment generally.

4.4 Results

In this section, I present the results of my investigation into the impact of adversarial techniques on user attitudes in conversational systems. I begin the analysis with an overview of the dependent variables, exploring their dynamics across three experimental conditions: Control, Need & Greed, and Time. I then provide a detailed discussion of each variable, i.e., trust, comfort level, willingness to disclose information, and perception of privacy and security. Data was rigorously analysed using repeated measures

ANOVA, Tukey's Post-Hoc tests, and boxplot visualisations to ensure accuracy. Statistical significance was established at $\alpha = 0.05$, indicating a 5% probability of a Type I error.

4.4.1 Overview of Dependent Variables

The dependent variables I investigated were chosen to understand user attitudes across the experimental conditions comprehensively. Trust, Comfort Level, Willingness to Disclose Information, and Perception of Privacy and Security were measured using targeted questions administered in post-task questionnaires after each experimental condition (Control, Need & Greed, Time). This approach provided valuable insights into how adversarial techniques influenced user attitudes.

Baseline Privacy Attitudes and Experimental Control

Baseline privacy attitudes were collected in the entry questionnaire to characterise the sample and provide context for interpreting findings. However, this study did not conduct formal statistical analyses using baseline privacy as a covariate (e.g., ANCOVA). Instead, the within-subject experimental design inherently controlled for individual differences in baseline privacy attitudes, as each participant served as their own control across the three experimental conditions.

This methodological approach is well-established in behavioural research [151]. By having participants experience all three conditions (Control, Need & Greed, Time) in a counterbalanced order, any influence of pre-existing privacy attitudes was distributed evenly across conditions. The repeated measures ANOVA used in this analysis automatically partitions out between-subject variance, effectively removing the influence of stable individual characteristics (including baseline privacy attitudes) from the error term when testing for condition effects.

To characterise the sample's baseline privacy profile, I examined responses to the entry questionnaire items measuring general attitudes toward online data sharing and prior experiences with conversational agents. Participants demonstrated variability

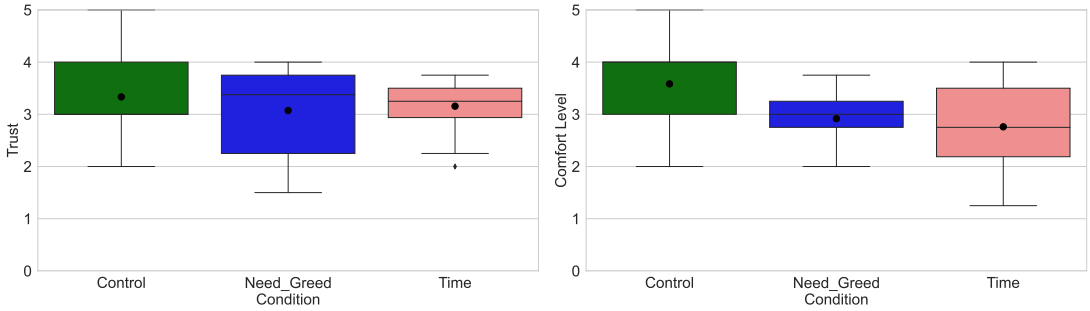
in their baseline privacy attitudes, with responses spanning the 5-point Likert scale, indicating that the sample included individuals with higher and lower levels of privacy consciousness. This diversity in baseline attitudes enhances the ecological validity of findings, as it reflects the natural variation found in real-world user populations.

Whilst additional exploratory analyses examining correlations between baseline privacy and post-task responses could have provided supplementary insights into individual differences, the primary research objective focused on changes in attitudes across experimental conditions (within-subject comparisons) rather than differences between individuals based on baseline characteristics (between-subject comparisons). The within-subject design, combined with Latin Square randomisation of condition order, provided robust experimental control for isolating the effects of adversarial techniques.

This approach represents a recognised methodological choice in experimental psychology, where within-subject designs are preferred when the research question centres on treatment effects rather than individual difference predictors. However, future research could benefit from explicitly modelling baseline privacy attitudes as a moderator variable to examine whether certain individuals are more or less susceptible to specific adversarial techniques.

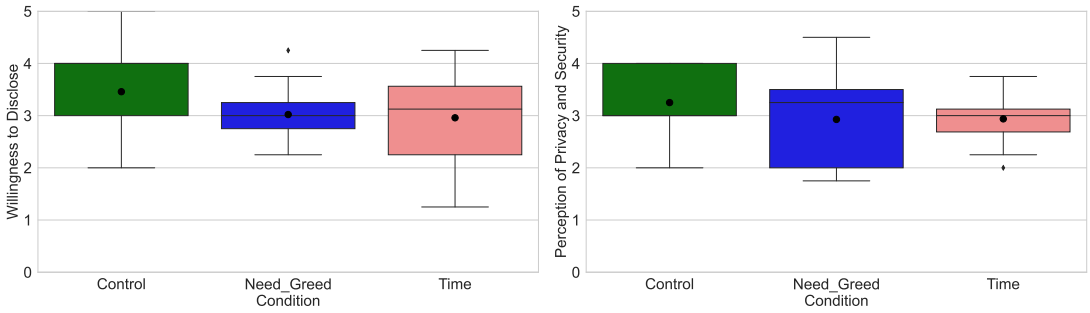
Table 4.2: Repeated measures ANOVA results for all dependent variables in the Wizard of Oz study (N=24). The table presents F values, degrees of freedom, and significance levels for Trust, Willingness to Disclose, Perception of Privacy and Security, and Comfort Level across three experimental conditions (Control, Need & Greed, Time). Significant effects (p¡0.05) were found for Willingness to Disclose (F=6.416, p=0.0035) and Comfort Level (F=8.9423, p=0.0005), whilst Trust and Perception of Privacy and Security showed no significant differences across conditions.

Dependent Variable	**F Value**	**Num DF**	**Den DF**	**Pr >F**
Trust	0.9474	2	46	0.3952
Willingness to Disclose	6.416	2	46	0.0035
Perception of Privacy and Security	1.8618	2	46	0.1669
Comfort Level	8.9423	2	46	0.0005



(a) Boxplot for Trust Scores across scenarios

(b) Boxplot for Comfort Level Scores



(c) Boxplot for Willingness to Disclose Information Scores across scenarios

(d) Boxplot for Perception of Privacy and Security Scores across scenarios

Figure 4.3: Boxplot distributions of participant responses (N=24) for four dependent variables measured on 5-point Likert scales across three experimental conditions. (a) Trust scores showing relatively consistent distributions across all conditions (F=0.9474, p=0.3952). (b) Comfort Level showing significant reductions in Need & Greed and Time conditions compared to Control (F=8.9423, p=0.0005). (c) Willingness to Disclose Information displaying lower median scores in Time condition (F=6.416, p=0.0035). (d) Perception of Privacy and Security with comparable distributions across conditions (F=1.8618, p=0.1669). Boxes represent interquartile ranges, horizontal lines indicate medians, and whiskers extend to 1.5 times the interquartile range.

4.4.2 Trust

I investigated participants' trust levels across the experimental scenarios to understand how adversarial techniques affect user trust in conversational systems. The ANOVA results ($F = 0.9474$, $p = 0.3952$) in Table 4.2 showed no statistically significant differences in trust between the Control, Need & Greed, and Time scenarios. The box plot (Figure 4.3a) further corroborated this consistency, with participants exhibiting comparable trust levels irrespective of the scenario. From these findings, I observed that trust remained stable even when adversarial techniques were introduced, suggesting that trust in conversational systems may be rooted in broader factors, such as users' general attitudes toward technology or prior experiences, rather than specific interaction conditions.

4.4.3 Comfort level

Comfort Level was significantly influenced by the experimental conditions ($F = 8.9423$, $p = 0.0005$) as shown in Table 4.2. Participants reported the highest comfort levels during the Control scenario, while both the Need & Greed and Time scenarios led to noticeable reductions, as illustrated in Figure 4.3b. Tukey's Post-Hoc Test in Table 4.3 confirmed these differences, with statistically significant drops in comfort when comparing the Control scenario to the other two. The results suggest that adversarial techniques, such as Need & Greed and Time, disrupted the natural flow of interaction, potentially adding stress or cognitive load. Reflecting on these findings, they underline the importance of designing conversational systems that prioritise user comfort, especially in scenarios requiring sensitive or high-stakes decision-making. Ensuring that users feel at ease is not just a matter of usability but also an ethical consideration.

4.4.4 Willingness to Disclose Information

The Willingness to Disclose Information showed significant variations across the experimental conditions ($F = 6.416$, $p = 0.0035$) as shown in Table 4.2. Tukey's Post-Hoc Test in Table 4.4) revealed that participants were significantly less willing to disclose

Table 4.3: Tukey's HSD post hoc pairwise comparisons for Comfort Level scores following significant ANOVA results (F=8.9423, p=0.0005). The table shows mean differences, adjusted p-values, 95% confidence intervals, and rejection decisions for all pairwise comparisons between experimental conditions. Significant differences were found between Control and Need & Greed (Mean Diff=-0.6667, p=0.0021) and between Control and Time (Mean Diff=-0.8229, p=0.0001), indicating that both adversarial techniques significantly reduced participant comfort compared to the Control condition.

Group 1	**Group 2**	**Mean Diff**	**p-adj**	**Lower**	**Upper**	**Reject**
Control	Need & Greed	-0.6667	0.0021	-1.1177	-0.2157	TRUE
Control	Time	-0.8229	0.0001	-1.2739	-0.3719	TRUE
Need & Greed	Time	-0.1562	0.6859	-0.6072	0.2947	FALSE

Table 4.4: Tukey's HSD post hoc pairwise comparisons for Willingness to Disclose Information following significant ANOVA results (F=6.416, p=0.0035). Only the comparison between Control and Time conditions reached statistical significance (Mean Diff=-0.5, p=0.0445), indicating that Time Pressure techniques specifically reduced participants' willingness to disclose information. The comparison between Control and Need & Greed approached but did not reach significance (p=0.0896).

Group 1	**Group 2**	**Mean Diff**	**p-adj**	**Lower**	**Upper**	**Reject**
Control	Need & Greed	-0.4375	0.0896	-0.9275	0.0525	FALSE
Control	Time	-0.5	0.0445	-0.99	-0.01	TRUE
Need & Greed	Time	-0.0625	0.9499	-0.5525	0.4275	FALSE

information in the Time scenario compared to the Control scenario ($p = 0.0445$). Differences between the Need & Greed scenario and the other conditions were not statistically significant.

The boxplot in Figure 4.3c highlighted the impact of time pressure, with the lowest willingness to disclose observed in the Time scenario. This suggests that participants adopt protective behaviours when they feel rushed or pressured. Reflecting on this, I think these findings reinforce the ethical responsibility to avoid designs that could manipulate users into sharing more information than they are comfortable with.

4.4.5 Perception of Privacy and Security

Perception of Privacy and Security did not show statistically significant differences across the experimental conditions ($F = 1.8618$, $p = 0.1669$) as shown in Table 4.2. However, the boxplot in Figure 4.3d indicated slight declines in privacy and security perceptions, particularly during the Time scenario. While the result did not reveal any significant effects, these minor trends are worth noting. This suggests that perceived urgency and pressure could subtly influence how users evaluate the safety of their data. Reflecting on these results, I believe further research, possibly involving different contexts, could uncover the nuanced ways in which adversarial conditions impact privacy and security perceptions. This additional insight would help design systems that better address user concerns and expectations.

4.4.6 Limitations and Mitigation Strategies

Whilst this study provides foundational insights into how adversarial techniques influence user attitudes in conversational systems, several limitations specific to this investigation must be acknowledged alongside the mitigation strategies employed to minimise their impact.

Wizard of Oz Methodology

The use of a human operator (wizard) to simulate system responses introduced potential concerns regarding operator bias and consistency across interactions. The wizard's interpretation of participant messages and selection of responses could have varied subtly between sessions, potentially introducing unintended variability.

Mitigation: I addressed these concerns through several mechanisms. First, I developed and used a pre-prepared library of standardised responses for each experimental condition, ensuring consistency in language, tone, and manipulation strategies across all participants. Second, I conducted two rounds of pilot testing (5 participants each) to refine the response protocols and identify technical issues before the main study. Pilot 1 revealed slow response times and technical link generation problems, both of which were resolved before Pilot 2. Pilot 2 confirmed that response timing felt natural and the experimental setup functioned correctly. Third, I strictly adhered to predefined response protocols throughout data collection, documenting any deviations for subsequent analysis. Finally, suspicion checks in exit questionnaires revealed that the vast majority of participants believed they were interacting with an autonomous system, confirming the effectiveness of the experimental setup.

Sample Characteristics

The sample was recruited through convenience sampling from academic networks (social media platforms and word-of-mouth within the University of Strathclyde community), resulting in a highly educated demographic (91.7% holding at least a bachelor's degree). This may limit the generalisability of findings to populations with different educational backgrounds or levels of technological familiarity.

Mitigation: Several measures were taken to address sampling concerns. First, the sample size (n=24) was determined through a priori power analysis (G*Power[4] 3.1, medium effect size f = 0.25, 80% power, $\alpha = 0.05$), ensuring sufficient statistical power

[4] G*Power: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologi
gpower

despite potential sample limitations. Second, the within-subject design reduced the impact of individual differences by having each participant serve as their own control, increasing statistical efficiency. Third, the sample exhibited diversity in age (18–54 years, M = 31.72, SD = 10.11) and gender (50% female, 50% male), providing some demographic variation. Fourth, baseline measures collected in the entry questionnaire enabled analysis of how individual differences influenced responses, allowing for more interpretable findings despite sample homogeneity in education level.

Order and Carryover Effects

The within-subject design, whilst offering advantages for statistical power, introduced the possibility that experiencing one experimental condition might influence responses to subsequent conditions. Participants' awareness of manipulation techniques in earlier scenarios could have increased vigilance in later interactions.

Mitigation: I implemented a Latin Square design to systematically randomise the presentation order of the three conditions (Control, Need & Greed, Time) across participants. This ensured that any order effects were distributed evenly across conditions and could not systematically bias the results. Additionally, condition-specific post-task questionnaires were administered immediately after each interaction, capturing attitudes before carryover effects could accumulate. The time gap between scenarios (during which participants completed questionnaires) also provided a cognitive break that helped reduce immediate carryover.

Ecological Validity

Participants were aware they were participating in research, which could have increased their vigilance and attentiveness to security concerns compared to real-world booking scenarios. Additionally, participants used dummy personal information (fictitious names, email addresses, phone numbers) during the task, which may have reduced the perceived risks associated with information disclosure.

Mitigation: Several design decisions enhanced ecological validity despite these con-

straints. First, the Y99 platform provided a familiar messaging interface similar to real-world applications (WhatsApp, Facebook Messenger), making interactions feel natural. Second, the flight booking task was designed with realistic parameters (route, dates, budget constraints) that required meaningful engagement with the conversational agent. Third, participants were encouraged to interact naturally and ask questions, mirroring typical user behaviour. Fourth, the web-based remote nature of the study allowed participants to engage from comfortable, familiar environments, reducing potential situational influences. Fifth, the use of dummy information, whilst protecting participant privacy, was standardised across all conditions, ensuring that any reduction in perceived risk affected all conditions equally and did not confound comparisons between conditions.

Single Domain and Limited Techniques

This study focused exclusively on a flight reservation scenario and examined only two adversarial techniques (Need & Greed and Time), limiting insights into how these techniques might operate across different domains or how other adversarial approaches might function.

Mitigation: This limitation was intentional at this foundational stage of the research. By focusing on a single, well-defined domain, I could establish baseline effects with tight experimental control before expanding to multiple domains in subsequent studies (chapters 5 and 6). The flight booking scenario was selected because it represents a common, relatable task where users regularly make decisions involving personal information and trade-offs. The methodological approach and findings from this study provided the foundation for the multi-domain investigation in chapter 5 and the expanded technique set (including Social Compliance) in chapter 6.

Baseline Privacy as a Covariate

Whilst baseline privacy attitudes were collected in the entry questionnaire, they were not formally analysed as covariates. The within-subject design inherently controlled

for individual differences by having participants serve as their own controls, effectively removing between-person variance in baseline attitudes from the error term. However, explicit analyses examining whether baseline privacy moderated the effects of adversarial techniques could have provided additional insights into individual susceptibility patterns.

Mitigation and Rationale: The decision to focus on within-subject changes across conditions rather than individual difference predictors was appropriate for the foundational research questions in this study, which centred on establishing whether adversarial techniques influence user attitudes rather than identifying who is most susceptible. The within-subject design with Latin Square counterbalancing provided robust control for individual differences. Nevertheless, this represents an area for future investigation, where baseline privacy attitudes could be explicitly modelled as moderating variables to identify characteristics of more vulnerable user populations. This limitation informed the design of subsequent studies (chapters 5 and 6), where additional baseline measures and more sophisticated analytical approaches were considered.

Despite these limitations and the mitigation strategies employed, this study's findings provide valuable foundational insights into how adversarial techniques influence user attitudes in conversational systems, establishing a baseline for the expanded investigations in subsequent chapters.

4.5 Summary

This chapter presented a foundational investigation into how adversarial techniques affect user attitudes in conversational systems, focusing specifically on Need & Greed and Time techniques. The research employed the Wizard of Oz framework to examine how these techniques influence trust, comfort levels, willingness to disclose information and perceptions of privacy and security.

section 4.3 then detailed the methodology, including the within-subject experimental design with 24 participants. It explained using the Wizard of Oz framework for simulating realistic interactions, developing three experimental conditions (Control,

Need & Greed, and Time), and the comprehensive data collection through pre-task, post-task, and exit questionnaires.

section 4.4 presented the findings through statistical analysis and visualisations. The results showed that while trust remained relatively stable across conditions, both Need & Greed and Time techniques significantly affected comfort levels and willingness to disclose information. The data revealed that adversarial techniques can disrupt the natural flow of interaction, potentially affecting how users make decisions about sharing information.

This study provides foundational evidence for the thesis statement in several ways. First, it demonstrates that adversarial techniques derived from scam principles (Need & Greed and Time) can indeed be systematically operationalised within text-based conversational systems through the Wizard of Oz framework. The significant effects observed on comfort levels ($F = 8.9423, p = 0.0005$) and willingness to disclose information ($F = 6.416, p = 0.0035$) validate the claim that these techniques meaningfully shape user attitudes. The stability of trust across conditions ($F = 0.9474, p = 0.3952$) whilst other attitudes changed significantly suggests that psychological vulnerabilities operate differently from what users consciously recognise as trustworthiness. These findings establish an empirical foundation for subsequent studies investigating how technique effectiveness varies across broader contexts and implementation technologies.

The findings of this chapter directly inform the expanded multi-domain investigation presented in chapter 5, particularly by highlighting the baseline behavioural responses to Need & Greed and Time Pressure techniques.

Chapter 5

Expanding the Horizons: contextual impacts of adversarial techniques in conversational systems

If my people, who are called by my name, will humble themselves and pray and seek my face and turn from their wicked ways, then I will hear from heaven, and I will forgive their sin and will heal their land. 2 Chron 7:14 (NIV)

Chapter 4 established baseline effects of Need & Greed and Time techniques in a single flight reservation context. This chapter extends the investigation across multiple domains (healthcare, online shopping, financial management, and home security) to examine how context shapes the effectiveness of adversarial techniques.

5.1 Introduction

The previous chapter demonstrated that Need & Greed and Time techniques significantly influence user comfort levels and willingness to disclose information in a controlled setting. However, the single-domain focus left open questions about how these effects generalise across different contexts with varying levels of sensitivity and user expectations.

Building on the findings from chapter 4, where I explored these techniques within a transactional context using a flight reservation task, this chapter broadens the scope to investigate their contextual dynamics across multiple domains. chapter 4 demonstrated

that Need & Greed and Time significantly affected comfort levels and willingness to disclose information, whereas trust and perceptions of privacy showed greater stability. However, the study was limited to a single domain, leaving unanswered questions about the generalisability of these effects across diverse real-world scenarios. This chapter addresses this gap by investigating how these techniques interact with contextual factors in domains such as healthcare, online shopping, financial management and home security.

I employed Twine, a narrative-driven tool for creating interactive scenarios to better understand the interplay between adversarial techniques and context. Twine allowed me to simulate realistic, domain-specific interactions that closely mirror real-world situations. For example, healthcare scenarios prompted users to share sensitive medical information under urgency, while online shopping scenarios leveraged promotional offers aligned with the Need & Greed principle. By tailoring scenarios to each domain, I aimed to investigate how adversarial techniques interact with context-specific dynamics. In addition to expanding the experimental scope, this chapter introduces Braun & Clarke's reflexive thematic analysis (TA) [172], a qualitative methodology suited to explore patterns in user attitudes.

Although chapter 4 relied primarily on quantitative analysis, the incorporation of TA provides a complementary lens, allowing for a deeper understanding of how adversarial techniques shape user attitudes across different contexts. Recent applications of thematic analysis in conversational systems research have demonstrated its value in capturing user perceptions and ethical concerns [179, 180]. By integrating this method, I aim to strengthen the research, offering insights that bridge quantitative trends with qualitative experiences.

5.2 Study Aim, Objectives, and Research Question

5.2.1 Study Aim.

The aim of this study, which operationalises dissertation-level Objective 2 (Section 1.4.2), is to examine how the adversarial techniques of Need & Greed and Time influence user attitudes across multiple domains, namely healthcare, online shopping, financial management, and home security, and how context-specific dynamics shape the effectiveness of these techniques in conversational systems.

5.2.2 Study Objectives.

The study-level objectives of this chapter, which operationalise dissertation-level Objective 2 (Section 1.4.2), are twofold.

1. To examine the impact of adversarial techniques across multiple domains, focusing on user attitudes such as trust, comfort levels, willingness to disclose information, and perceptions of privacy and security.

2. To explore how contextual factors shape the influence of adversarial techniques, revealing domain-specific variations and their implications for ethical design.

5.2.3 Study Research Questions.

The following study-level research questions, which refine dissertation-level RQ2 (Section 1.4.3), guide these objectives:

RQ 5.1: How do the adversarial techniques of Need & Greed and Time influence user attitudes across different scenarios in conversational systems?

RQ 5.2: How does the interaction between different scenarios and the adversarial techniques shape user attitudes in conversational systems?

Key findings from this chapter include the following.

- Both techniques significantly increased comfort levels and willingness to disclose information across all contexts tested.
- Trust levels exhibited context-specific variations, with scenarios involving high personal or financial stakes showing greater sensitivity.
- Perceptions of privacy and security were relatively stable, suggesting that factors beyond immediate interaction conditions may shape these attitudes.
- Context-specific dynamics highlighted the need for domain-aware system designs that address unique ethical challenges.

Combining Twine and thematic analysis, this chapter builds on the foundations established in chapter 4 to provide a detailed, context-sensitive understanding of adversarial techniques in conversational systems. The findings contribute to developing ethical, user-centred conversational systems, offering actionable insights for designers and researchers. These insights emphasise the importance of balancing user engagement with transparency and safeguarding autonomy while highlighting areas for further exploration, such as domain-specific implications and long-term effects on user attitudes.

5.2.4 Contributions

The contributions of this chapter are as follows.

1. A systematic investigation of how adversarial techniques influence user attitudes across healthcare, online shopping, financial management, and home security domains.
2. Advancement in methodology from Wizard of Oz to the Twine platform, enabling more controlled and scalable investigation of user responses across multiple scenarios.
3. Integration of reflexive thematic analysis with quantitative measures to provide deeper insights into how users respond to adversarial techniques in different contexts.

4. Experimental evidence of how different adversarial techniques manifest in LLM-based interactions, thus establishing a foundation for future research.

5.3 Methodology

This section outlines the detailed steps I undertook to investigate how adversarial techniques, i.e. Need & Greed and Time, affect user attitudes across different contexts. I designed the study to ensure the scenarios were as realistic as possible, mirroring the types of interactions users might encounter in real-world applications. My aim was to capture both the immediate and contextual effects of these techniques while ensuring that participants could engage meaningfully with the study tasks.

5.3.1 Experimental Design

The experimental setup was designed to combine interactive, narrative-driven scenarios with structured data collection tools (questionnaire sets). Using Twine, an open-source tool, I developed branching, domain-specific scenarios that mirrored real-world interactions across healthcare, online shopping, financial management, and home security. This allowed me to examine how the adversarial techniques of Need & Greed and Time influenced participants' attitudes within different contexts studied.

The scenarios in Appendix A.1, such as healthcare, online shopping, financial management, and home security, were designed to ensure participants could engage naturally while embedding adversarial techniques in ways reflective of standard practices in real-world systems. Scenarios began with a brief introduction to set the context, after which participants interacted with a conversational system tailored to the specific domain. For instance,

- In the online shopping scenario, Need & Greed was operationalised through incentives such as discounts or exclusive offers to encourage information sharing.
- In the home security scenario, urgency was introduced through "limited-time offers" tied to security system upgrades.

This approach allowed me to simulate authentic interactions while ensuring that the effects of the adversarial techniques were observable across participants' trust, comfort levels, willingness to disclose information and perceptions of privacy and security.

Rationale for Method Selection

The selection of Twine as the implementation platform for this study addressed specific limitations identified in the WoZ study (chapter 4). The WoZ framework required real-time human operation for each session, constraining scalability and introducing potential operator fatigue. Additionally, the single-domain focus necessitated examining adversarial techniques across multiple contexts.

Twine offered three key advantages for this research stage. First, its scripted passages ensured standardised implementation, eliminating subtle response variations inherent in human-operated systems. Second, the modular structure enabled efficient development of scenarios across four domains (healthcare, shopping, financial, home security) whilst maintaining consistent technique operationalisation. Third, web-based delivery allowed participant autonomy and self-paced interaction, enhancing ecological validity.

Twine's limitations (scripted interactions and inability to handle free-text input) were acceptable for this research stage, providing the methodological bridge between controlled WoZ implementation and the naturalistic LLM-based study in chapter 6.

Pilot Testing

Two pilot rounds were conducted before the main study to refine the Twine-based scenarios, test the experimental protocol, and identify potential technical or usability issues. The pilot tests were essential for ensuring that the branching narrative structure functioned correctly across all four domains (healthcare, online shopping, financial management, and home security) and that participants could navigate the scenarios intuitively.

Pilot 1: The initial pilot involved 5 participants recruited through Prolific using

the same eligibility criteria as the main study (age 18+, English fluency, approval rate 80–100%). This pilot revealed two significant issues that required attention. First, several participants reported confusion about the scenario randomisation mechanism, as the transitions between domains felt abrupt without adequate contextual framing. Second, the timing of post-task questionnaire presentation was inconsistent, with some participants being redirected to Qualtrics before completing all scenario branches.

Pilot 2: A second pilot round with five different participants tested the revised implementation. All participants completed the procedure successfully, and the scenario transitions were perceived as natural and contextually appropriate. The Qualtrics integration functioned correctly also. Participants' qualitative feedback indicated that the scenarios felt authentic and that the adversarial techniques were not overtly obvious, confirming that the experimental manipulations were appropriately subtle. Post-pilot debriefing revealed that participants found the dummy personal information (section A.2) easy to use and that the overall study flow felt coherent. This confirmed that the experimental setup, scenario design, questionnaire timing, and data capture procedures were ready for the main study. Participants from both pilot rounds were excluded from the main study to prevent contamination effects.

Information Request Consistency

To isolate the effects of adversarial techniques from information sensitivity, the types of information requested remained consistent across all three experimental conditions (Control, Need & Greed, Time) within each scenario. For example, in the Healthcare scenario, all conditions requested basic health information (age, weight, existing medical conditions) and potentially sensitive data (Medical Insurance number or NHS details). The key manipulation was **how** this information was requested (through neutral dialogue in Control, incentive framing in Need & Greed, or urgency framing in Time) rather than **what** information was requested. This design ensured that observed differences in user attitudes (trust, comfort, willingness to disclose, privacy concerns) could be attributed to the adversarial techniques themselves rather than to differential information sensitivity across conditions. The dummy information sheet (Appendix

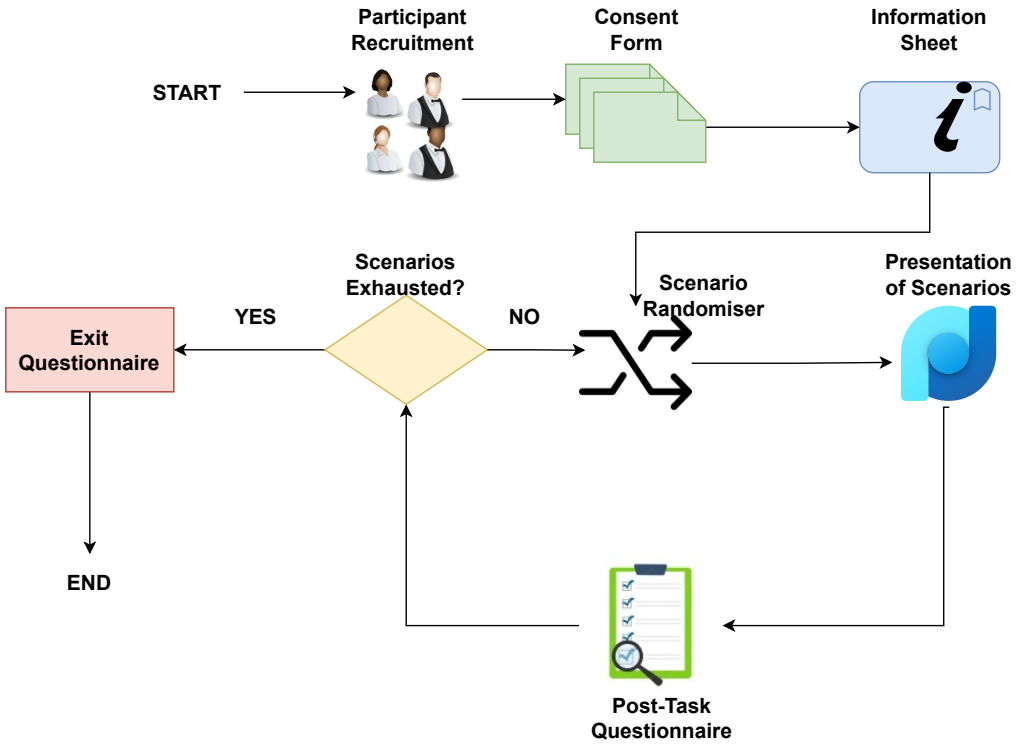


Figure 5.1: Experimental procedure flowchart for the Twine-based multi-domain study. The diagram illustrates the complete participant journey from initial recruitment through Prolific Academic, consent procedures, information sheet review, randomised scenario presentation across four domains (Healthcare, Online Shopping, Financial Management, Home Security), post-task questionnaires administered via Qualtrics after each scenario, and culminating in an exit questionnaire. The decision point ensures participants complete all scenarios before final submission, with scenarios exhausted indicated by the binary decision node.

A.2) provided standardised data across all scenarios, further controlling for disclosure risk perception. By maintaining consistent information requests whilst varying only the conversational framing, the experimental design enabled clear attribution of effects to the specific adversarial techniques under investigation.

5.3.2 Procedure

As shown in Figure 5.1, the experimental procedure illustrates the progression from participant recruitment to debriefing.

Sample Size Determination

I conducted an a priori power analysis using G*Power 3.1.9.7 to determine the appropriate sample size for this study. The analysis was configured for repeated measures ANOVA with a within-factors design, reflecting the study's structure, where each participant experienced all combinations of conditions and scenarios.

Based on previous research examining user attitudes in conversational systems and the effect sizes observed in the WoZ study (chapter 4), I anticipated a medium effect size (f = 0.25). The power analysis used the following parameters: effect size f = 0.25, α level of 0.05, power = 0.90, number of groups = 1, number of measurements = 12 (representing the 4 domains $\times$ 3 techniques matrix), correlation among repeated measures = 0.5, and nonsphericity correction epsilon = 1.

The analysis indicated that a minimum of 16 participants would be required to achieve 90% power for detecting medium-sized effects. To account for potential participant attrition and to enhance statistical robustness, I recruited 30 participants, nearly double the minimum required sample size. This provided an actual power of 0.92, ensuring sufficient sensitivity to detect the effects of adversarial techniques across multiple domains whilst maintaining adequate statistical power even if some participant data required exclusion.

Participant Recruitment and Compensation

Participants were recruited using the Prolific platform, ensuring a diverse and qualified pool. I targeted individuals aged 18 or older with an approval rate of 80 to 100%, indicating reliability in previous studies. Fluency in English was required due to the study's language needs. I excluded individuals who had participated in the study's previous iterations (pilots).

Upon successful study completion, each participant received £10 compensation through the Prolific platform. Compensation was contingent on data completeness, with all submissions reviewed to ensure participants had completed all required scenarios and questionnaires. This fixed-rate compensation model ensured fairness across all

participants and was determined based on Prolific's recommended payment guidelines for the estimated study duration (approximately 40–60 minutes). Payment was processed through Prolific's secure payment system within 48 hours of study completion and approval.

This recruitment strategy resulted in the successful enrolment of $n = 30$ approved participants, comprising 14 males and 16 females, aged 23 to 60 years ($M = 37.5$, $SD = 10.18$). Detailed demographic characteristics are presented in Table 5.1.

Table 5.1: Demographic characteristics of participants in the Twine-based multi-domain study (N=30). The table presents gender distribution (46.7% male, 53.3% female), age ranges (23–60 years, M=37.5, SD=10.18), and education levels. The sample was recruited through Prolific Academic with an approval rate of 80–100%, ensuring participant reliability. The majority (43.3%) held undergraduate degrees, with educational attainment ranging from secondary education to graduate degrees.

Characteristic	**Category**	**N**	**%**
Gender	Male	14	46.7
	Female	16	53.3
Age (years)	23–30	9	30.0
	31–40	8	26.7
	41–50	10	33.3
	51–60	3	10.0
	Mean (SD)	37.5 (10.18)	
Education	Secondary education	1	3.3
	High school/A-levels	7	23.3
	Technical/community college	5	16.7
	Undergraduate degree	13	43.3
	Graduate degree	4	13.3
Total		**30**	**100**

Study workflow

Once participants were selected, they received an information sheet detailing the study's objectives, procedures, and ethical safeguards, including data anonymity and their right to withdraw at any time. Before proceeding, participants had to electronically sign a consent form acknowledging their understanding of the study and their voluntary

participation (see B.2.1).

Participants began the study by accessing the Twine scenarios through a randomised scenario selector, ensuring no fixed order of domains. After each scenario, they completed a post-task questionnaire assessing their trust, comfort levels, willingness to disclose information, and perceptions of privacy and security. Dummy data (see Appendix A.2), such as pre-generated name and email address, was provided to maintain participant privacy and standardise the risk of information sharing. Once all scenarios were completed, the participants completed an exit questionnaire to provide feedback on their overall experience and perceptions of adversarial techniques. The complete post-task questionnaires for the Control, Need & Greed, and Time conditions are provided in Appendices B.2.2, B.2.3 and B.2.4 respectively. The exit questionnaire is provided in Appendix B.2.5. This design enabled me to collect quantitative and qualitative data to understand how Need & Greed and Time techniques influence user attitudes.

5.4 Result

In this section, I present the findings of my investigation into the effects of adversarial techniques, specifically Need & Greed and Time, on user attitudes across the four scenarios used. My analysis is guided by the two research questions (RQ 5.1 and RQ 5.2) involving both quantitative and qualitative analysis. The quantitative analysis focuses on statistical results from repeated measures ANOVA and Tukey's post hoc tests to compare responses across treatments and scenarios. I also used the violin plots to summarise key trends in trust, comfort level, willingness to disclose information, and perceptions of privacy and security. The qualitative analysis complements the statistical findings by exploring patterns and themes in participants' experiences and responses, providing deeper insight into how scenarios and adversarial techniques influenced their attitudes.

5.4.1 Quantitative Analysis

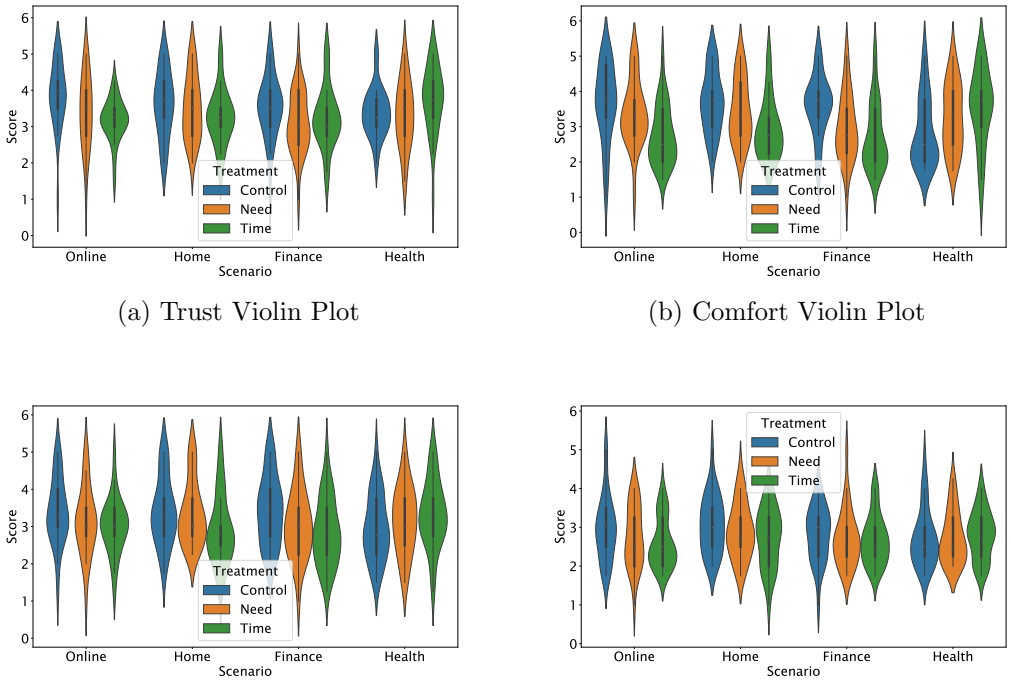
The data analysis began by evaluating the responses from 30 participants across all experimental conditions. Data from a single participant were excluded due to incompleteness, leaving 29 valid responses for detailed analysis. The study employed a 5-point Likert scale to capture participants' responses under each scenario, with the results processed through repeated measures ANOVA to examine the main effects and interactions. The results for each dependent variable are detailed below, accompanied by statistical tables and plots to illustrate the findings.

Trust

Trust was evaluated to understand how adversarial techniques and scenarios influenced participants' confidence in the conversational system under different conditions. Trust was measured using a set of four 5-point Likert scale questions in each experimental condition (Control, Need & Greed, and Time Principle). While the core trust construct remained consistent, the questions were adapted to reflect each condition's specific context. For example, in the Control condition, participants were asked: "How much did you trust the conversational agent's advice or suggestions in this interaction?" (1: Not at all - 5: Completely), while in the Time Principle condition, they rated "how the time-sensitive nature of the scenario impacted your trust in the agent's advice" (1: No impact - 5: High impact). This approach allowed for comparative analysis of trust across different experimental conditions.

As shown in Table 5.2, the repeated measures ANOVA revealed a significant main effect of Treatment ($F(2, 56) = 4.40$, $p = 0.0168$) and a significant Scenario $\times$ Treatment interaction ($F(6, 168) = 4.14$, $p < 0.001$). However, the main effect of Scenario alone was not significant ($F(3, 84) = 1.73$, $p = 0.168$), indicating that variations in trust were primarily driven by the manipulation type and its interaction with contextual domain rather than scenario alone.

Figure 5.2a shows that trust scores were broadly distributed across all conditions. Under the Need & Greed treatment, trust scores decreased noticeably in the financial



(a) Trust Violin Plot

(b) Comfort Violin Plot

(c) Willingness Violin Plot

(d) Security Violin Plot

Figure 5.2: Violin plots displaying the distribution and density of participant responses (N=29, after one exclusion) across four domains (Online Shopping, Home Security, Financial Management, Healthcare) for three experimental conditions (Control, Need & Greed, Time). Each subfigure shows: (a) Trust scores measured on a 5-point Likert scale, with broader distributions in Healthcare and Financial domains; (b) Comfort Level scores showing narrower distributions in Control conditions compared to adversarial techniques; (c) Willingness to Disclose Information with notably lower scores in Time conditions across most domains; (d) Perception of Privacy and Security showing relatively consistent distributions across conditions.

management scenario, reflecting participants' scepticism toward monetary incentives. In contrast, the healthcare scenario under the Time treatment showed slightly higher trust scores, suggesting that urgency might foster trust in time-sensitive situations. These findings suggest that while trust is generally resilient to scenario-specific influences, adversarial techniques like Need & Greed and Time can significantly shape how users perceive the reliability of conversational agents. For instance, urgency might heighten trust in healthcare scenarios, where speed is often associated with competence [53].

Table 5.2: Repeated measures ANOVA results for trust levels across treatments and scenarios. Significant effects were observed for Treatment and the Scenario × Treatment interaction, while Scenario alone was non-significant, suggesting that user trust varied as a function of manipulation type and its contextual application.

Effect	**F Value**	**Num DF**	**Den DF**	**Pr >F**
Scenario	1.728	3	84	0.1675
Treatment	4.400	2	56	0.0168*
Scenario × Treatment	4.137	6	168	0.0007***

Further analysing the significant interaction effects for Scenario × Treatment ($p < 0.001$) on trust, Tukey's post hoc test was applied to identify which conditions differed significantly. The results summarised in Appendix A.3 revealed patterns in how participants' trust varied based on context. Trust levels under the Control condition were significantly higher in financial management and healthcare scenarios compared to online shopping ($diff = 1.0517$ and 1.10172 respectively, $p < 0.01$), likely reflecting people's inherent trust in established institutions. This makes practical sense as financial institutions and healthcare providers typically operate under strict regulations and professional standards. This naturally inspires more trust than in online shopping environments where vendors might be unknown or less regulated. Interestingly, this pattern was reversed in the Need & Greed condition.

Trust levels dropped more in the financial management and healthcare scenarios compared to online shopping. This suggests that when people perceive attempts to manipulate their needs or desires, they become particularly suspicious. This heightened suspicion likely reflects the higher stakes involved in making decisions in these contexts or scenarios.. However, trust drops less in online shopping, possibly because people already expect some level of pressure in such scenarios. Under the Time condition, trust levels remained high, especially in the financial management scenario compared to online shopping ($diff = 0.9914$, $p = 0.0082$). This suggests that the Time-based adversarial technique may be less effective because people are more accustomed to making time-pressured financial decisions. However, in the healthcare scenario, the lower trust observed under time pressure may reflect people's preference for unhurried

medical decisions.

Comfort Level

Comfort levels were assessed to measure participants' perceived ease or readiness to proceed with decisions under each condition. This was measured using four 5-point Likert scale questions across all three experimental conditions (Control, Need & Greed, and Time Principle), with items tailored to each. For example, in the Control condition, participants rated "How comfortable did you feel during your interaction with the conversational agent?" (1: Very uncomfortable - 5: Very comfortable), while in the Need & Greed condition, they assessed "Rate your comfort level when the conversational agent offers specific incentives" using the same scale. This allowed for evaluating how different adversarial techniques affected participant comfort during interactions.

The repeated measures ANOVA results shown in Table 5.3 revealed significant main effects for Treatment ($F(2, 56) = 10.53$, $p = 0.0001$) and a significant Scenario $\times$ Treatment interaction ($F(6, 168) = 9.26$, $p < 0.0001$). However, the main effect of Scenario alone was not significant ($F(3, 84) = 1.75$, $p = 0.164$), indicating that variations in participants' comfort levels were primarily driven by manipulation type and its contextual interaction rather than scenario alone.

Table 5.3: Repeated measures ANOVA results for comfort levels across treatments and scenarios. Significant effects were found for Treatment and for the Scenario $\times$ Treatment interaction, while Scenario alone was non-significant, suggesting that user comfort depended on both manipulation type and its contextual application rather than scenario independently.

Effect	**F Value**	**Num DF**	**Den DF**	**Pr >F**
Scenario	1.746	3	84	0.1639
Treatment	10.534	2	56	0.0001***
Scenario $\times$ Treatment	9.255	6	168	<0.0001***

Figure 5.2b shows that comfort levels were highest under control conditions and dropped significantly under the Need & Greed treatment, particularly in financial management scenarios. Conversely, the Time treatment elicited higher comfort levels in healthcare

scenarios, possibly due to the relevance of urgency in such contexts. These results indicate that adversarial techniques disrupt the natural flow of interaction, with Need & Greed negatively impacting comfort levels. The findings highlight the importance of designing conversational systems that minimise stress, especially in scenarios requiring sensitive decision-making.

To further investigate the significant interaction effect for Scenario × Treatment ($p < 0.0001$) on comfort levels, Tukey's post hoc analysis, as summarised in Appendix A.3, indicated that under control conditions the participants' comfort levels were significantly higher in financial management and healthcare scenarios compared to home security and online shopping scenarios ($diff = 0.9569$, $p = 0.0154$). This pattern is consistent with real-world behaviours where engaging with banking apps or healthcare portals is routine. In contrast, making decisions about home security can feel more personal and stressful as it directly affects private spaces and family safety.

Under the Need & Greed condition, this pattern changed. While healthcare scenarios maintained relatively high comfort levels, there was a decline in the financial management scenario ($diff = -0.9483$, $p = 0.0180$). This is understandable as individuals are more accepting of doctor-recommended health interventions but may become uncomfortable when pushed to make financial decisions. Health needs are perceived as more genuine or time-sensitive, whereas money-related pressure tactics may trigger a natural sense of distrust.

A different pattern emerged when time pressure was introduced under the Time condition. The drop in comfort levels was more in the online shopping scenario ($diff = -0.931$, $p = 0.0244$). By contrast, healthcare and financial management scenarios exhibited greater resilience to time pressure, likely because people are somewhat accustomed to making quick judgments in these regulated environments (e.g., urgent medical advice or routine banking operations). Institutions tend to project enough credibility in these contexts that participants remain relatively comfortable with rapid decision-making. However, online shopping environments with many unknown vendors appear more vulnerable to decreases in comfort levels when users sense urgency.

Willingness to Disclose Information

Willingness to disclose information was analysed to understand participants' openness to sharing personal or sensitive data. The ANOVA results (Table 5.4) showed significant effects for Treatment ($p = 0.0148$) and Scenario $\times$ Treatment ($p = 0.0007$), while the Scenario effect ($p = 0.6615$) was not significant.

Figure 5.2c reveals that willingness to disclose information decreased under the Need & Greed treatment, particularly in financial management scenarios, where participants were likely to be more protective of sensitive data. Conversely, time treatment in healthcare scenarios showed an increased willingness to disclose, suggesting that urgency might reduce reluctance to share information. These findings highlight the ethical implications of using adversarial techniques to influence user attitudes and behaviour. While urgency can encourage disclosure in critical scenarios, such as healthcare, it risks being manipulative if applied indiscriminately.

Further analysing the interaction effect for Scenario $\times$ Treatment, Tukey's Post Hoc test, as summarised in Appendix A.3, indicated that under control condition, both healthcare and home security scenarios showed a higher willingness to disclose information compared to online shopping scenario ($diff = 0.8966$, $p = 0.0433$ for both). This aligns with real-world behaviours in two ways: participants recognise that accurate medical care often depends on complete and honest disclosure in healthcare contexts. Strong privacy laws and ethical standards also help reduce concerns over sharing sensitive health data. Under home security scenarios, sharing personal details with security providers makes one feel more justified regarding protecting one's home and family.

Under the Need & Greed condition, a different pattern emerged. While the healthcare scenario revealed the highest willingness to disclose, the financial management scenario showed the steepest drop ($diff = 0.9483$, $p = 0.0180$, suggesting that in healthcare, individuals remain relatively open to sharing information due to clear, tangible health benefits and the perceived legitimacy of why these data are requested. In financial management scenarios, pressure tactics raise concerns about the misuse

of sensitive information, as money-related decisions are often approached with greater caution. These response patterns may vary significantly across different cultural contexts, as cultural norms, institutional trust levels, and varying privacy expectations shape attitudes toward healthcare and financial information disclosure. Future cross-cultural studies would be valuable in determining how these factors affect vulnerability to adversarial techniques in different regions.

Under the time condition, healthcare scenarios showed an increased willingness to disclose information, further highlighting the influence of urgency. In situations perceived as time-sensitive, individuals may feel an immediate need to cooperate, particularly when health or safety is on the line. However, this same urgency could be ethically problematic if deployed in less critical contexts, as it may nudge users to reveal more than they otherwise would under normal conditions.

The repeated measures ANOVA results shown in Table 5.4 revealed a significant main effect of Treatment ($F(2, 56) = 4.55$, $p = 0.0148$) and a significant Scenario $\times$ Treatment interaction ($F(6, 168) = 4.11$, $p = 0.0007$). However, the main effect of Scenario alone was not significant ($F(3, 84) = 0.53$, $p = 0.662$), indicating that participants' willingness to disclose information varied primarily as a function of manipulation type and its interaction with the contextual domain.

Table 5.4: Repeated measures ANOVA results for willingness to disclose information. Significant effects were found for Treatment and the Scenario $\times$ Treatment interaction, while Scenario alone was non-significant, indicating that disclosure willingness depended on manipulation type and its contextual application rather than scenario alone.

Effect	**F Value**	**Num DF**	**Den DF**	**Pr >F**
Scenario	0.532	3	84	0.6615
Treatment	4.549	2	56	0.0148*
Scenario $\times$ Treatment	4.106	6	168	0.0007***

Perceived Privacy and Security

Perceived privacy and security were analysed to examine how adversarial techniques influenced participants' confidence in the system's ability to safeguard the information they provided. The repeated measures ANOVA results (Table 5.5) revealed a marginally significant main effect of Treatment ($F(2, 56) = 3.24$, $p = 0.0466$), while both Scenario ($F(3, 84) = 1.45$, $p = 0.234$) and the Scenario × Treatment interaction ($F(6, 168) = 1.33$, $p = 0.246$) were not significant. These results suggest that participants' perceptions of privacy and security were modestly affected by manipulation type but remained largely stable across scenarios.

Table 5.5: Repeated measures ANOVA results for perceived privacy and security. A marginally significant main effect was found for Treatment, whereas Scenario and the Scenario × Treatment interaction were non-significant, indicating that perceived privacy and security were only modestly influenced by manipulation type and unaffected by context.

Effect	F Value	Num DF	Den DF	Pr >F
Scenario	1.452	3	84	0.2337
Treatment	3.240	2	56	0.0466*
Scenario × Treatment	1.330	6	168	0.2463

This suggests that perceived privacy and security are relatively stable across scenarios but can be slightly influenced by adversarial techniques. The non-significant scenario effect contrasts with patterns observed in other measures (e.g., trust, comfort level, willingness to disclose information), where scenario-specific differences were clear. In this case, participants seem to have relatively fixed views about system security, likely formed from their prior experiences and general attitudes toward data protection.

The effect of Treatment on privacy and security perceptions was marginal, implying that adversarial techniques can still nudge users' sense of safety. However, the change appears to be limited. When these techniques are applied, they may prompt suspicion. However, these shifts do not appear as pronounced as in measures discussed previously.

5.4.2 Qualitative Analysis

To gain a deeper understanding of the complex interactions between user experiences, psychological factors, and behavioural responses to conversational systems, I employed Braun and Clarke's thematic analysis (TA) method. This approach allowed me to identify, analyse systematically, and report patterns within the qualitative data obtained from the exit interviews. Following Braun and Clarke's approach, the reflexive thematic analysis process involved six key phases: familiarisation with the data, generating initial codes, searching for themes, reviewing potential themes, defining and naming themes, and finally, producing the report [172]. Each phase was conducted carefully to ensure the reliability and depth of our qualitative insights. Through thematic analysis, I uncovered seven key themes directly linked to our research questions and objectives. These themes provide a comprehensive view of user attitudes and behaviour in response to the adversarial techniques employed in conversational systems. The complete codebook with definitions, inclusion and exclusion criteria, and example quotes is provided in Appendix A.4.

I will now describe how I implemented the six phases before presenting the findings.

Phase 1 - Familiarisation with the data

None, however I am very sceptical of online Ai or chatbots and those not being upfront about what is done with my data (P1)

I have had concerns about entering card details (P13)

I felt that the responses and the site design didn't look secure so that put me off (P13)

I have been scammed before so felt wary when card details were asked for (P8)

Some of these felt like chat gippity and some felt like click funnels. But none of them were giving me good vibes. (P4)

I've had many scam calls regarding phones asking for private information off the bat and its always very pressing. This is very off putting so any agent which asks for private information quickly is usually a scam. (P15)

Figure 5.3: Phase 1 of Braun and Clarke's reflexive thematic analysis showing the immersion and familiarisation process with qualitative data from post-task questionnaire responses (N=29 participants). The diagram illustrates the initial coding framework with main categories including Privacy and Security Concerns, Decision-Making Processes, Agent Characteristics, Psychological Triggers, Past Digital Experiences, and Technology Interaction Alignment. This phase involved repeated reading of participant responses to identify preliminary patterns and areas of interest for subsequent detailed coding.

The goal of this phase is to thoroughly immerse in the data by reading and re-reading it, taking note of initial thoughts, potential patterns, and emerging themes. This process helps build a deep understanding of the content before moving on to the next phase. I reviewed each participant's response multiple times to grasp the overall content. A sample of responses is shown in Figure 5.3. I began to notice recurrent topics and concerns by engaging deeply with the data. Through this initial familiarisation, I recognised the importance of context, especially participants' previous experiences with scams and their heightened security concerns. This stage laid the foundation for the coding process by establishing a clear understanding of the data.

Phase 2 - Generating Initial Codes.

The purpose of this phase is to code the data systematically. This involves identifying significant features across the entire dataset and coding them in a consistent manner. Each code represents a key feature of the data, helping to break down the data into manageable pieces for further analysis. During the coding process, I applied specific labels to key excerpts from the dataset that seemed relevant to the research questions, as shown in Table 5.6. These codes were guided by the familiarisation process and aligned with the initial patterns and emerging themes. These initial codes were broad and aimed at capturing all the relevant details from the participants' responses. This comprehensive approach allowed for the identification of patterns across different participants' experiences and ensured that no significant aspect of the data was overlooked. During the coding process, I applied specific labels to key excerpts from the dataset that seemed relevant to the research questions, as shown in Table 5.6. These codes were guided by the familiarisation process and aligned with the initial patterns and emerging themes. These initial codes were broad and aimed at capturing all the relevant details from the participants' responses. This comprehensive approach allowed for the identification of patterns across different participants' experiences and ensured that no significant aspect of the data was overlooked. The final codebook, including detailed definitions and application criteria for each code, is provided in Appendix A.4.

Ensuring Codebook Reliability and Consistency

Following established guidelines for rigorous qualitative research [181], I implemented several measures to ensure the reliability and consistency of the coding process. These measures were essential for establishing the trustworthiness of the thematic analysis and ensuring that the codes and themes accurately represented the participants' experiences.

Codebook Development and Refinement. The initial codebook was developed through an iterative process during Phase 2. I began by independently coding a subset

Excerpt	Code
"I have been scammed before so felt wary when card details were asked for" (P8)	Past Digital Experiences
"I 've had many scam calls regarding phones asking for private information off the bat and its always very pressing. This is very off putting so any agent which asks for private information quickly is usually a scam." (P15)	Past Digital Experiences
"I try to keep very secure so I withhold any information possible." (P15)	Privacy and Security Practices
"I'm reluctant to give away information or data easily without a specific reason" (P1)	Privacy and Security Practices
"I disliked the hurried high pressure sales tactics " (P1)	Agent Characteristics
" The non pushy and caring tone made me feel safer than the pushy time sensitive ones." (P3)	Agent Characteristics
"I was thinking how much I trusted the conversational agent , and whether I could trust them with my sensitive information" (P1)	Decision-making Processes
"I was thinking about whether the interaction seemed genuine ,how much information was being asked for and whether the scenario justified the amount of information requested." (P23)	Decision-making Processes
" Thinking you have a bargain when actually you don't need it " (P13)	Psychological Triggers
"I felt a sense of insecurity when reading about the high instance of burglaries in my area and this made me feel motivated to purchase security systems." (P23)	Psychological Triggers
" The urgency made me rush in my decisions " (P16)	Time Pressure
" Time constraints and pressure immediately turn me off from engaging . This is often listed as a red flag for fraud." (P29)	Time Pressure

Table 5.6: Phase 2: Generating Initial Codes from the Data.

of five participant responses (approximately 17% of the dataset) to generate preliminary codes. These initial codes were refined through multiple rounds of review, where I examined whether each code had clear definitional boundaries, was applied consistently, and captured meaningful aspects of the data. Code definitions were documented with inclusion and exclusion criteria, along with exemplar quotes that illustrated each code's application. For instance, the code "Past Digital Experiences" was defined as references to prior interactions with digital platforms, scams, or security breaches that shaped current attitudes, whilst excluding general statements about technology use without specific experiential grounding.

Systematic Application and Consistency Checks. To ensure consistent application of codes across the entire dataset, I employed a systematic approach. After establishing the initial codebook, I coded all 29 participant responses whilst maintaining detailed analytical memos that documented coding decisions, ambiguities, and emerging patterns. I then conducted consistency checks by revisiting previously coded segments after completing the full dataset to verify that codes remained stable and were applied uniformly. Instances where coding decisions were ambiguous were flagged and revisited, with decisions documented in the audit trail. This iterative checking process ensured that codes maintained consistent meaning throughout the analysis.

Reflexivity and Researcher Positionality. Throughout the coding process, I maintained reflexivity by acknowledging my position as a researcher investigating adversarial techniques and recognising how this perspective might influence code interpretation. To mitigate potential bias, I deliberately sought contradictory evidence and alternative interpretations during coding. For example, when identifying instances of privacy concerns, I explicitly looked for and coded cases where participants expressed minimal concern or trust in the system, ensuring balanced representation of the data. Regular reflexive journaling documented my analytical decisions, assumptions, and evolving interpretations, creating transparency in the research process.

Peer Debriefing and Verification. To enhance the credibility of the coding framework, I engaged in peer debriefing sessions with my doctoral supervisor, who has expertise in human-computer interaction and qualitative methodology. During these sessions, we reviewed a sample of coded data (approximately 20% of responses), discussed coding decisions, and examined whether codes were appropriately applied and themes adequately supported by the data. This process provided an external check on the analytical process and helped identify areas where code definitions required clarification or refinement. Feedback from these sessions led to minor adjustments in code boundaries and more precise theme definitions.

Audit Trail and Transparency. I maintained a detailed audit trail throughout the analysis, documenting the evolution of codes, theme development decisions, and analytical rationale. This included preserving earlier versions of the codebook, noting when and why codes were merged or split, and recording the progression from initial codes to final themes. This documentation ensures transparency and allows for verification of the analytical process. Table 5.6 presents the final codebook with representative excerpts, demonstrating how codes were applied to actual participant data.

These combined measures (systematic codebook development, consistency checks, reflexivity, peer debriefing, and transparent documentation) ensured that the coding process was reliable and that the resulting themes accurately represented participants' experiences across the dataset. Whilst Braun and Clarke's reflexive thematic analysis does not require inter-rater reliability statistics (as the approach values researcher subjectivity and interpretation), these quality measures demonstrate methodological rigour and enhance the trustworthiness of the findings [181, 179].

Phase 3 - Searching for Themes

In this phase, the goal was to sort the various codes into themes and to collate all relevant data extracts within each theme. The themes were broader than the initial codes and were intended to capture the overall patterns in the data, reflecting the participants' collective experiences and concerns. I began grouping similar codes based

on the coded data to form overarching themes. For example, all codes related to past negative experiences with digital platforms were grouped under the theme Past Digital Experiences. Similarly, concerns and practices related to privacy and security were combined into Privacy and Security Practices and Privacy and Security Concerns. This phase involved examining the relationships between the codes and how they connected to form a coherent story about the participants' experiences (Figure 5.4). I aimed to ensure the themes were meaningful and distinct, each capturing a significant aspect of the data.

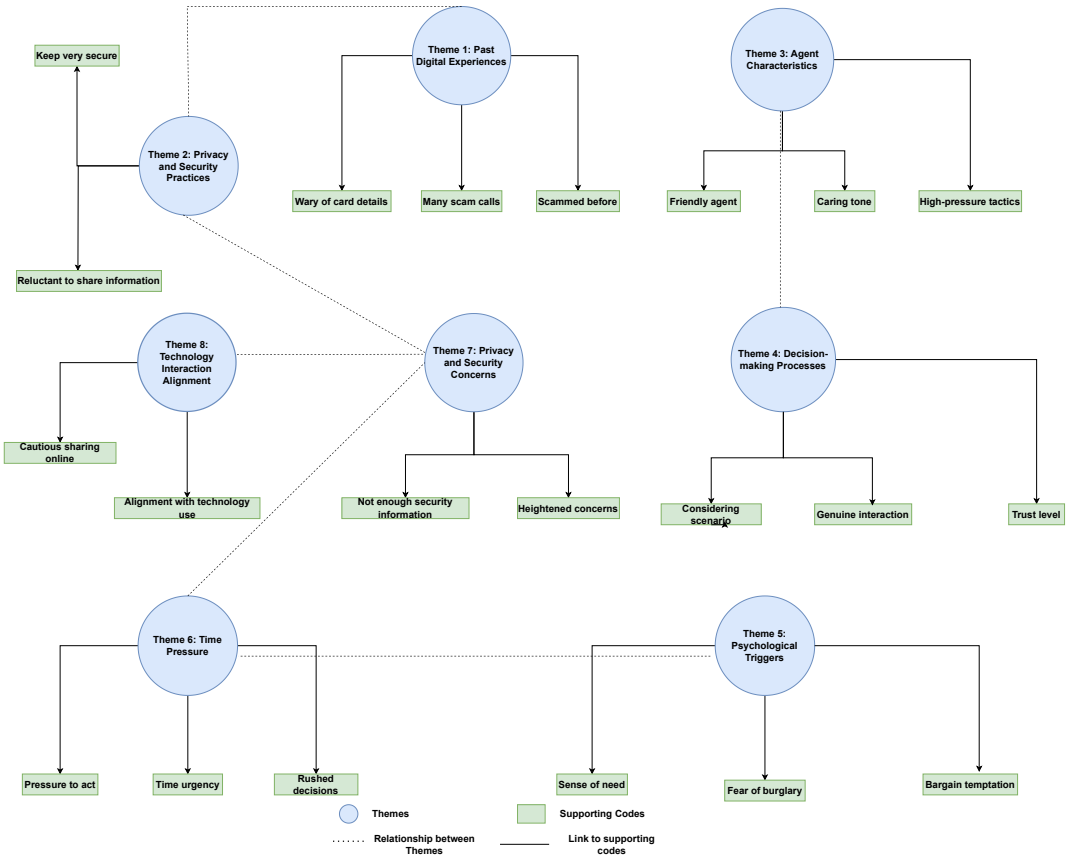


Figure 5.4: Phase 3 of reflexive thematic analysis illustrating the thematic map with seven preliminary themes: Privacy and Security Concerns, Privacy and Security Practices, Agent Characteristics, Decision-Making Processes, Psychological Triggers, Time Pressure, and Technology Interaction Alignment. Connecting lines represent relationships between themes, such as Agent Characteristics influencing Decision-Making Processes. This phase involved grouping codes into themes and examining their interrelations to form a coherent narrative of participant experiences.

Phase 4 - Reviewing Themes

This phase involves refining the themes to reflect the coded data and broader dataset accurately. The aim is to review each theme, verify its relevance, and check if it works in relation to the data. During this phase, I revisited the coded data extracts for each theme to ensure they fit well together and accurately represented the data. I also checked the themes against the entire dataset to ensure no significant data was overlooked. Initially, the Psychological Triggers theme was broad and included motivations such as need, greed, and fear. Upon review, it became evident that the theme was too diverse. To address this, I refined it by splitting the triggers into more specific categories, such as Need and Fear-based Triggers. Privacy and security concerns were initially coded into two themes: Privacy and Security Practices and Privacy and Security Concerns. However, these two themes overlapped significantly, as both dealt with participants' cautious behaviour regarding their data. Therefore, I merged them into a Privacy and Security Considerations theme to streamline the analysis.

Reviewed Themes

- Theme 1: Past Digital Experiences remained unchanged as the data consistently supported it.
- Theme 2: Privacy and Security Considerations emerged from the merging of Privacy and Security Practices and Privacy and Security Concerns, as both dealt with how participants navigated privacy and security issues based on their past experiences and concerns.
- Theme 3: Agent Characteristics continued to be supported by data highlighting how the conversational agent's tone, responsiveness, and demeanour influenced trust and comfort levels.
- Theme 4: Decision-making processes remained a distinct theme, capturing the thought processes participants engaged in during key moments of interaction, particularly regarding trust and information disclosure.

- Theme 5: Psychological triggers were refined into Need-based and Fear-based triggers, though these subthemes were later reintegrated after reviewing their interconnections.

- Theme 6: Time Pressure was validated as a significant standalone theme, with multiple participants noting the impact of urgency on their decisions.

- Theme 7: Technology Interaction Alignment was maintained to encapsulate how participants' choices during the experiment aligned with their general approach to technology, including their comfort levels with conversational systems and digital platforms.

After this review, the themes were clear and coherent and captured the essential aspects of participants' experiences. This refinement process ensured that each theme was supported by robust data and reflected the underlying patterns in participants' responses.

Phase 5 - Defining and Naming Themes

This phase aims to clearly define each theme and ensure that each tells a coherent story about the data. This involves finalising theme names, developing a narrative for each theme, and ensuring that each theme contributes to answering the research questions. During this phase, I created detailed definitions for each theme, ensuring they accurately captured the essence of the coded data. The goal was to articulate the scope of each theme, describe its content, and determine how it related to the overall research questions.

Final Themes and Definitions

Theme 1: Past Digital Experiences: This theme encapsulates participants' prior experiences with digital platforms, particularly those involving scams or distrust, and how these shaped their cautious behaviour in the experiment. To illustrate this theme, I present a direct quote from a participant:

"I have been scammed before so felt wary when card details were asked for" (P8)

Theme 2: Privacy and Security Considerations: This theme includes all aspects of participants' concerns and practices around privacy and security, showing how these considerations influenced their decisions regarding trust and information disclosure.

"I'm uncomfortable sharing personal information online" (P12)

Theme 3: Agent Characteristics: This theme focuses on how the conversational system's tone, responsiveness, and behaviour impact participants' trust and comfort levels.

"The non-pushy and caring tone made me feel safer" (P3)

Theme 4: Decision-making Processes: This theme covers the cognitive processes participants engaged in when deciding whether to trust the agent, disclose information, or make purchasing decisions.

"I was considering whether the interaction seemed genuine" (P23)

Theme 5: Psychological Triggers: This theme captures the influence of psychological triggers, such as need, fear, and greed, on participants' decisions.

"Thinking you have a bargain when actually you don't need it" (P13)

Theme 6: Time Pressure: This theme highlights how the urgency created by the conversational agent influenced participants' trust and decision-making processes.

"The urgency made me rush in my decisions" (P16)

Theme 7: Technology Interaction Alignment: This theme explores how participants' general attitudes toward technology use and their comfort with AI aligned with their choices in the experiment.

"The conversational agent is taking over but they try too hard to be an over-friendly human and that freaks me out" (P20)

The theme names were chosen to be descriptive and reflect the data they represented (Figure 5.5). Each name was designed to immediately convey the core idea of the theme to ensure clarity and focus during the final analysis.

Phase 6 - Producing the Report

The final phase involves writing the analysis and presenting the themes in a coherent, persuasive narrative that answers the research questions. This phase includes linking the themes to the broader literature and drawing out implications for practice or further research.

Each theme is presented in detail, supported by relevant data extracts. I connected the themes to existing research on user behaviour with digital platforms and conversational agents, highlighting how this study contributes new insights, particularly regarding the psychological triggers and decision-making processes influenced by agent characteristics and time pressure.
I diligently maintained reflexivity throughout the thematic analysis process, ensuring any potential bias and assumption were acknowledged and addressed. This approach was crucial in enhancing the credibility and validity of my findings.

Theme 1: Past Digital Experiences. Participants' experiences with digital platforms significantly shape their attitudes towards current interactions. Past negative experiences, such as scams or privacy breaches, are mentioned as factors that led to increased scepticism. An example is: *"I have been scammed before, so I felt wary when card details were asked for" (P8).* Interestingly, these accounts of prior scams and wariness echo the quantitative finding that willingness to disclose information was notably lower in financial management scenarios under the Need & Greed condition ($p = 0.0180$). This suggests that negative past experiences amplify users' distrust, specifically in money-related contexts.

This finding aligns with research on trust in online environments, suggesting that trust is fragile and easily broken by negative experiences [182]. Users can carry this lack of trust forward in digital contexts, applying it to new interactions [50]. The emphasis

on past experiences highlights the importance of first impressions and the long-term impact of the trustworthiness of digital service providers. This observation is in sync with the [183] trustworthiness model, which underscores the role of perceived benevolence and integrity in forming trust judgements. Participants' reluctance to fully engage with the conversational agent can be a protective mechanism, a concept supported by the 'privacy calculus model' where users weigh the benefits against potential privacy risks [47].

Theme 2: Privacy and Security Considerations This theme is pervasive throughout the responses. Participants are cautious, often withholding information or requiring a solid justification for data sharing. An example is: *"I'm reluctant to give away information or data easily without a specific reason" (P1).* This caution aligns with our quantitative results: perceived privacy and security were modestly but significantly influenced by adversarial treatments ($p = 0.0466$), while willingness to disclose information also dropped notably under Need & Greed ($p = 0.0148$). Such findings reinforce how heightened privacy worries can directly translate into reduced data-sharing.
This behaviour reflects a risk-averse attitude toward privacy, consistent with previous research showing that privacy concerns often motivate protective behaviours in online environments [145]. Participants' emphasis on cautious information disclosure also reflects the privacy paradox, in which stated privacy concerns do not always correspond with actual behaviour. This careful approach aligns with explanations offered by information boundary theory, which proposes that individuals manage their privacy through boundary regulation based on perceived risks and benefits [184]. The reluctance to share information, therefore, can be understood as a boundary management strategy intended to protect personal data from potential misuse.

Theme 3: Agent Characteristics The responses indicate that characteristics such as tone, responsiveness, and presentation of information by the conversational agent significantly affect decision-making. A friendly, non-pressuring tone fosters trust, whereas a pushy or hurried approach triggers discomfort and retreat. An example is: *"I*

disliked the hurried high-pressure sales tactics." (P1). Quantitatively, this resonates with earlier findings that comfort levels declined substantially under the Need & Greed treatment ($p = 0.0001$), and trust was significantly shaped by the scenario × treatment interaction ($p < 0.001$). In other words, pushy conversational systems can erode user comfort and trust, as reflected in the statistical results.

These observations align with earlier research showing that a computer's perceived personality can shape the way users interact with it [185]. The influence of tone in user–agent communication is particularly relevant within the framework of Media Equation theory, which proposes that individuals often respond to computers and other media as if they were human [186]. Studies have further shown that the tone and behaviour of an agent can activate social heuristics, leading users to apply social norms and expectations to technological interfaces [187]. The persuasive technology framework also illustrates how such agents can guide or influence user actions, reinforcing the importance of designing interactions that respect user comfort and autonomy.

Theme 4: Decision-making processes. The participants' decision-making processes reveal a complex interplay of trust, comfort, and information relevance. Decisions about trust and information disclosure are not made lightly; rather, they result from careful consideration of the agent's trustworthiness and the perceived value of the interaction. An example is: *"I was considering whether the interaction seemed genuine" (P23).* In line with the quantitative results, these detailed decision-making deliberations mirror the significant scenario × treatment interactions observed for both trust ($p < 0.001$) and willingness to disclose information ($p = 0.0007$).

Participants' real-time judgments about the conversational system's genuineness and the context at hand strongly influenced their choices. This resonates with the theory of reasoned action [188], which explains how individuals make reasoned decisions based on attitudes and social norms. The role of cognitive heuristics in decision-making is also highlighted, aligning with the work on judgment under uncertainty [112]. Participants seem to apply heuristics such as 'representativeness' and 'availability' when evaluating conversational agents, which affects their willingness to engage.

Theme 5: Psychological Triggers. The influence of psychological triggers, such as a perceived scarcity or fear of missing out (FOMO), is critical. Participants report a heightened sense of caution when they perceive manipulative tactics, which could be interpreted as a trigger for psychological reactance. This motivational state occurs in response to threats to perceived behavioural freedoms. An example is: *"Thinking you have a bargain when actually you don't need it" (P13).* Such responses map onto the earlier quantitative observation that comfort levels dropped significantly under the Need & Greed prompts ($p < 0.0001$), suggesting that when users sense overt manipulative strategies like dangling bargains, they may respond with increased resistance rather than compliance.

This reactance may explain participants' resistance to high-pressure tactics as they seek to reassert their autonomy. These findings align with the principles of persuasion [28], particularly scarcity, which can create a sense of urgency. However, the responses in this study suggest that such tactics can backfire in digital environments, especially when users are wary of manipulative intent.

Theme 6: Time Pressure. Time pressure has a significant impact on how individuals interact with technology, often producing feelings of discomfort and distrust. Participants generally prefer to make decisions at their own pace and respond negatively when external time limits are imposed. This pattern of behaviour is consistent with findings from concept analyses of information overload, which describe it as a negative psychological state in which individuals feel overwhelmed by excessive information, reducing their capacity to perform tasks effectively [189]. Such research indicates that time pressure can produce both emotional effects, including stress and anxiety, and cognitive effects, such as diminished attention and difficulties in processing information. This helps explain why the Time technique observed in this study triggered strong discomfort responses: It deliberately creates conditions that promote information overload through artificial urgency, reducing users' cognitive capacity for critical evaluation. An example is: *"Time constraints and pressure immediately turn me off from engaging" (P29).* Correspondingly, the quantitative data revealed a significant drop in comfort

levels under Time-based adversarial conditions ($p < 0.0001$), especially in the online shopping scenario, thereby highlighting that rushed tactics can quickly erode user ease and receptivity.

This aligns with Trope and Liberman's temporal construal concept, where time pressure can distort information processing and decision-making [29]. The elaboration likelihood model suggests that under time constraints, individuals can resort to the peripheral processing of persuasive messages, leading to superficial consideration and potential resistance [190]. This highlights the importance of allowing users ample time to process the information for a more thoughtful engagement.

Theme 7: Technology Interaction Alignment. The degree to which participants' interactions with conversational agents align with their general approach to technology is crucial in their acceptance of technology. The participants' dispositions and experiences highlight that accepting a technology system is not just about the system's features but also about individual differences in technology use and attitudes. An example is: *"The conversational agent is taking over, but they try too hard to be an over-friendly human." (P20).* This sentiment aligns with the moderate yet significant main effect of adversarial treatments ($p = 0.0466$) on perceived privacy and security, highlighting how a mismatch between user expectations and the behaviour of the conversational system can weaken comfort, trust and general acceptance of the system.

This finding echoes the technology acceptance model, suggesting that perceived ease of use and usefulness are key determinants of technology adoption. The concept of cognitive fit [191] also resonates here, indicating that successful interaction with technology depends on alignment with the established mental models of users and comfort with the use of technology.

5.5 Limitations

Whilst this study advances understanding of adversarial techniques across multiple contexts, several limitations specific to the Twine-based implementation and design choices warrant acknowledgement.

5.5.1 Platform and Interaction Constraints

The use of Twine for scenario implementation, whilst enabling standardised and scalable delivery across four domains, introduced inherent constraints in interaction naturalness. Participants navigated pre-scripted branching pathways rather than engaging in free-text conversations, limiting the authenticity of the conversational experience. This constraint meant that adversarial techniques were experienced through predetermined dialogue options rather than emerging organically through dynamic exchanges.

Mitigation and Forward Path: Several design features minimised this limitation's impact. The branching narrative structure provided multiple pathways based on participant choices, creating agency despite the scripted foundation. Dummy personal information (Appendix A.2) standardised disclosure risk whilst maintaining privacy. The four distinct domains were carefully designed to mirror realistic scenarios, enhancing ecological validity. Post-task and exit questionnaires included open-ended questions allowing participants to articulate experiences beyond the constrained interactions. Importantly, this limitation directly motivated the progression to the LLM-based study (chapter 6), which addressed these constraints through fully dynamic conversational interactions.

5.5.2 Within-Subject Design Considerations

The within-subject design, whilst statistically efficient, introduced potential carryover effects as participants experienced all twelve scenario-condition combinations (4 domains $\times$ 3 conditions). Repeated exposure might have sensitised participants to adversarial techniques or led to fatigue effects.

Mitigation: Latin Square randomisation systematically counterbalanced presentation order, distributing any carryover effects evenly rather than systematically biasing particular conditions. Domain randomisation further controlled for learning effects. Statistical analyses revealed no significant order effects, suggesting effective counterbalancing.

5.5.3 Measurement Approach

Attitudes were measured exclusively through self-report questionnaires administered after scenario completion rather than capturing behavioural data during interactions. This reliance on post-hoc self-reports introduced potential for social desirability bias and memory effects, whilst missing opportunities to observe actual disclosure behaviours.

Mitigation and Forward Path: Several measures enhanced validity. Anonymous Prolific-based data collection reduced social desirability pressures. Dummy information eliminated genuine disclosure risk, enabling authentic responses. The short duration between interaction and measurement (immediate post-task questionnaires) minimised memory decay. Thematic analysis revealed consistency between quantitative patterns and qualitative narratives, suggesting self-reports captured genuine attitudes. This limitation informed the LLM study (chapter 6), which incorporated behavioural metrics (actual information disclosure, response patterns, resistance strategies) captured during real-time interactions.

5.5.4 Sample and Generalisation

The sample (N=30) was recruited through Prolific with inclusion criteria requiring English fluency and prior experience with online systems. The resulting sample was predominantly Western, relatively educated (50% undergraduate, 20% graduate degrees), and likely more technologically literate than the general population.

Mitigation: Power analysis confirmed adequate sensitivity (92% power) for detecting medium-sized effects. The within-subject design maximised statistical efficiency.

Demographic diversity in age (23–60 years) and gender (53% female) provided variation. Thematic analysis revealed heterogeneity in past digital experiences and privacy attitudes, suggesting meaningful diversity in technological perspectives despite geographical and educational similarities. Future research should prioritise international samples to examine cross-cultural variations.

5.6 Summary

This chapter extended the investigation of adversarial techniques in conversational systems, demonstrating how Need & Greed and Time techniques can influence user attitudes across four distinct domains. I combined quantitative analysis with qualitative insights through Braun & Clarke's thematic analysis to understand how these techniques affect user trust, comfort levels, willingness to share information and perceived privacy and security in varied scenarios.

First, section 5.1 introduced the research context, explaining how this study builds upon chapter 4's findings. It outlined the use of Twine for creating interactive scenarios and established the research questions guiding the investigation. section 5.3 then detailed the methodology, describing the experimental design using Prolific for participant recruitment and implementing four distinct scenarios: healthcare, online shopping, financial management, and home security. It explained how the study gathered quantitative and qualitative data through post-task questionnaires and exit interviews.

section 5.4 presented the findings through two main approaches. The quantitative analysis revealed significant effects of adversarial techniques on user attitudes, with variations across different contexts. The qualitative analysis, using Braun & Clarke's thematic approach, identified seven key themes: past digital experiences, privacy and security considerations, agent characteristics, decision-making processes, psychological triggers, time pressure, and technology interaction alignment that helped explain how users respond to adversarial techniques in different settings.

section 5.5 acknowledged the study's limitations whilst outlining mitigation strategies. Key limitations included Twine's scripted interaction constraints, potential within-

subject carryover effects, reliance on self-report measures, and sample characteristics. Each limitation was addressed through specific design decisions, including branching narratives for enhanced realism, Latin Square randomisation for counterbalancing, immediate post-task measurement, and adequate statistical power. These limitations and their mitigation strategies directly informed methodological refinements in subsequent research.

The findings of this multi-domain investigation directly support the thesis claim that adversarial technique effectiveness varies significantly based on context and domain. The significant Scenario $\times$ Treatment interactions observed for trust ($p = 0.0001$), comfort ($p < 0.0001$), and willingness to disclose ($p = 0.0007$) empirically validate that psychological vulnerabilities manifest differently across healthcare, financial management, online shopping, and home security contexts. Healthcare scenarios demonstrated distinctive resistance patterns to Need & Greed techniques but heightened responsiveness to Time pressure, whilst financial contexts showed the highest overall vulnerability to Need & Greed manipulations. This context-dependent variation confirms that user attitudes and behaviours are shaped not solely by the techniques themselves but by the interplay between manipulation strategy and domain sensitivity. These findings reinforce the thesis assertion that psychological vulnerabilities must be recognised alongside technical security measures in conversational system design, as the same technique can elicit vastly different user responses depending on the perceived sensitivity and legitimacy of the interaction context.

Insights gained from context-specific vulnerabilities identified here inform the subsequent LLM-based investigation in chapter 6, shaping both methodological refinements and highlighting areas for deeper exploration of user compliance and resistance behaviours.

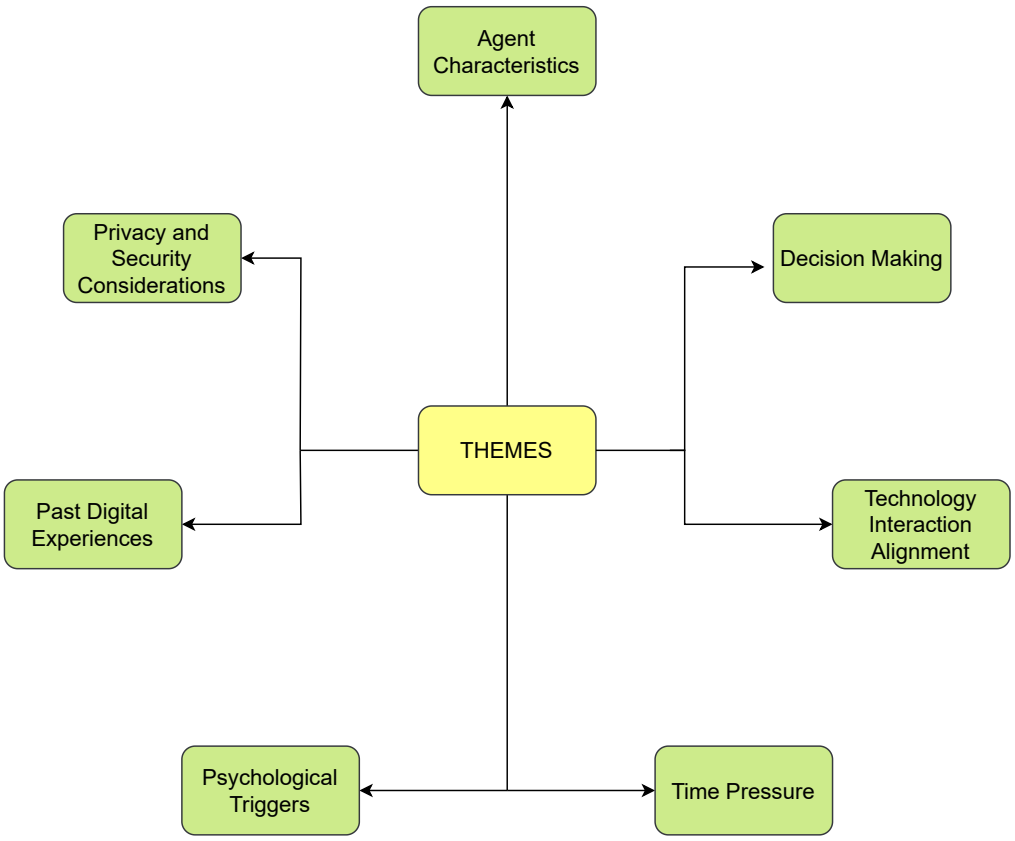


Figure 5.5: Final thematic map showing refined themes and their hierarchical relationships following Phase 5 of Braun and Clarke's reflexive thematic analysis. The map presents three overarching themes—Privacy and Security Awareness, Influence Recognition and Response, and System Interaction Dynamics—each with related subthemes. Line thickness represents the relative strength of relationships between themes, highlighting how participants experienced and responded to adversarial techniques across contexts.

Chapter 6

Investigating Social Compliance and Need & Greed Adversarial Techniques in LLM-Based Conversational Systems

Trust in the Lord with all your heart and lean not on your own understanding. Prov 3:5 (NIV)

Chapters 4 and 5 examined adversarial techniques in controlled (Wizard of Oz) and narrative-based (Twine) environments. This chapter investigates how Social Compliance and Need & Greed techniques manifest in dynamic, LLM-based conversational systems powered by the Llama API, where real-time interaction and adaptive responses create new opportunities for manipulation.

6.1 Introduction

The progression from human-operated simulations (chapter 4) to interactive narratives (chapter 5) established that adversarial techniques significantly influence user attitudes across contexts. This chapter advances the investigation by examining how these techniques operate in LLM-based systems that generate contextually relevant responses in real time. For example, threat actors exploit GenAI in prompt injection attacks, inputting carefully crafted prompts to manipulate AI systems [192] or create adversarial

examples that generate hostile traffic targeting intrusion detection systems [193]. Building on the findings from chapter 4 and chapter 5, which explored adversarial techniques in controlled and multi-domain settings, this chapter investigates their implementation within LLM-based conversational systems. LLMs, known for their dual utility in defensive and adversarial applications, have a profound impact on user interactions with search interfaces and cybersecurity measures. Their ability to generate human-like, highly convincing content enhances phishing attacks [194], introduces vulnerabilities in suggested codes that sophisticated attackers can exploit [195], and facilitates Cyber-enabled Social Influence Operations (CeSIOs). In such operations, LLMs are deployed to craft deceptive online identities and engage authentic users, subverting discourse and shaping public opinion on social media platforms [196].

This chapter investigates how Social Compliance and Need & Greed adversarial techniques manifest and operate within LLM-based conversational systems, and examines their influence on user behaviour and attitudes during real-time interactions with AI-driven dialogue systems. By studying these techniques in dynamic LLM environments, I can better understand how their implementation differs from static conversational agents and what new security implications might arise.

Rationale for Technique Selection The decision to replace Time pressure techniques with Social Compliance in this chapter was informed by both methodological considerations and the progressive nature of my research. Having investigated Time techniques across two previous studies (Chapters 4 and 5), I had already established how time techniques influence user attitudes in conversational systems. The findings demonstrated that Time techniques create immediate psychological effects, particularly in contexts where users perceive urgency as legitimate (such as healthcare scenarios in chapter 4). This understanding provided a sufficient empirical foundation regarding temporal manipulation in conversational systems.

An important practical consideration influenced the decision to examine only two techniques in this study. Implementing all three techniques (Need & Greed, Time, and Social Compliance) simultaneously would have substantially increased participant fa-

tigue. With three experimental conditions plus a control condition, participants would have needed to complete four full conversational scenarios, each followed by comprehensive questionnaires measuring trust, comfort, willingness to disclose information, privacy concerns, and security perceptions. This extended protocol would have introduced significant fatigue effects, potentially compromising data quality and participant engagement, particularly given the cognitively demanding nature of conversational interactions with LLM-based systems. By limiting the investigation to two adversarial techniques alongside the control condition, I maintained methodological rigour whilst ensuring reasonable participation duration and meaningful engagement throughout the experiment.

The introduction of Social Compliance techniques in chapter 6 served several objectives. First, it allowed for the examination of a distinct psychological mechanism, one rooted in authority, social proof, and obedience principles rather than temporal urgency. This broadened the scope of adversarial techniques investigated across the thesis whilst maintaining focus on Need & Greed as a consistent comparison point across all three studies.

Second, Social Compliance techniques proved particularly well-suited to the dynamic capabilities of LLM-based systems. Unlike the Time technique, which relies primarily on creating urgency through explicit temporal constraints, Social Compliance operates through more subtle conversational mechanisms: establishing authority through linguistic markers, demonstrating expertise through contextual knowledge, and leveraging social norms through references to typical user behaviour. These elements align naturally with the generative capabilities of LLMs, which can dynamically construct authoritative language patterns and contextually appropriate demonstrations of expertise in ways that static or semi-scripted systems cannot.

Third, the methodological progression from controlled (Wizard of Oz) through semi-structured (Twine) to dynamic (LLM) implementations presented an opportunity to investigate how different adversarial techniques manifest across varying levels of system sophistication. Examining Social Compliance, specifically in its most sophisticated implementation, provided insights into how authority-based manipulation operates in

systems capable of generating highly contextual, responsive content. Finally, the combination of Need & Greed and Social Compliance in chapter 6 provided an empirically valuable comparison between reward-based and authority-based compliance mechanisms within the same dynamic conversational environment, offering insights that complemented the reward-based versus urgency-based comparison from the earlier studies.

6.2 Study Aim, Objectives, and Research Question

6.2.1 Study Aim.

The aim of this study, which operationalises dissertation-level Objectives 3 and 4 (Section 1.4.2), is to investigate how Social Compliance and Need & Greed adversarial techniques manifest in dynamic LLM-based conversational systems, examining their influence on user behaviour and attitudes during real-time interactions, and how users' defensive behaviours and resistance strategies evolve across scenarios over time.

6.2.2 Study Objectives.

The study-level objectives of this chapter, which operationalise dissertation-level Objectives 3 and 4 (Section 1.4.2), are as follows:

1. To examine how Social Compliance and Need & Greed adversarial techniques influence user behaviour and attitudes within LLM-based conversational systems implemented through a Flask application calling the Llama-3.3-70B API.

2. To investigate how contextual factors (scenario type) moderate the effectiveness of these techniques in dynamic LLM-mediated dialogue.

3. To analyse how users' defensive behaviours and resistance strategies develop and evolve during sustained interactions with adversarial LLM-based systems.

6.2.3 Study Research Questions.

The following study-level research questions, which operationalise dissertation-level RQ3 and RQ4 (Section 1.4.3) within an LLM-based implementation, guide this study:

RQ 6.1: How do Social Compliance and Need & Greed adversarial techniques influence user behaviour and attitudes in LLM-based Conversational systems?

RQ 6.2: How do contextual factors (scenario type) moderate the effectiveness of adversarial techniques in LLM-based conversational systems?

RQ 6.3: How do users' defensive behaviours evolve during interactions with different adversarial techniques in LLM-based systems?

RQ 6.4: How do users develop and modify their resistance strategies across different scenarios and adversarial techniques in LLM-based conversational systems?

6.2.4 Contributions

The contributions of this chapter are as follows.

1. A systematic investigation of how Social Compliance and Need & Greed adversarial techniques operate within dynamic LLM-based conversational systems
2. Analysis of real-time user responses and defensive adaptations to adversarial techniques across multiple scenarios
3. Implementation of a novel experimental framework that integrates LLM capabilities for studying adversarial techniques in conversational systems
4. Evidence-based guidelines for designing secure LLM-based conversational systems that protect against social engineering and information disclosure risks

The rest of this chapter is organised as follows.

section 6.3 details the experimental methodology, including system implementation and scenario design. section 6.4 presents the findings and their implications for secure

conversational systems. section 6.6 concludes with key insights and directions for future research.

6.3 Methodology

This section describes the steps that I took to investigate how adversarial techniques, specifically Social Compliance and Need & Greed, affect user attitudes and behaviours in LLM-based conversational systems. I designed the experimental framework to comprehensively evaluate these techniques under realistic conditions. The objective was to capture immediate behavioural responses and longer-term attitudinal changes while ensuring that participants could engage meaningfully with the system.

6.3.1 Methodological Justification

The selection of a Flask-based web application integrated with Llama 3.3-70b addressed fundamental limitations of previous methodologies whilst aligning with the specific research objectives of this investigation.

Addressing Limitations of Previous Methodologies The Wizard of Oz framework employed in chapter 4, whilst providing experimental control, relied on human-operated responses that limited scalability and introduced potential operator variability despite standardised protocols. The Twine-based platform in chapter 5, though effective for examining context-specific vulnerabilities, operated through predetermined conversational pathways. This scripted nature constrained the investigation of how users respond to adaptive, contextually aware adversarial techniques. Users navigated fixed decision trees rather than engaging with systems capable of generating novel, contextually appropriate responses, particularly limiting the examination of defensive behaviours and resistance strategies. The research questions guiding this chapter required a methodology capable of capturing real-time user responses, defensive adaptations, and evolving resistance strategies. These objectives necessitated a system generating contextually appropriate responses to diverse user inputs whilst maintaining experi-

mental control over adversarial technique implementation. LLM-based systems offered the technological sophistication necessary to examine these dynamics.

Flask Framework Selection Flask's lightweight architecture provided the flexibility needed to integrate the Llama[1] API whilst maintaining fine-grained control over experimental conditions. Its routing capabilities enabled distinct experimental conditions (Control, Need & Greed, Social Compliance) within a unified application architecture, ensuring consistent technical infrastructure whilst allowing systematic variation through system messages. The framework's Python compatibility facilitated integration with data analysis pipelines, enabling real-time logging of user interactions, response latencies, and compliance patterns.

Llama 3.3-70b Selection Llama 3.3-70b was selected for several reasons. First, it represented a state-of-the-art open source model with capabilities comparable to proprietary alternatives, ensuring the research examined adversarial techniques in systems reflecting current technological sophistication. Second, its 70 billion parameter architecture provided sufficient capacity for generating contextually rich responses across diverse scenarios without extensive fine-tuning, essential for implementing Social Compliance techniques requiring domain specific expertise and authority markers. Third, its open source nature enabled transparent deployment whilst facilitating replicability. Fourth, pilot testing confirmed the model could maintain coherent multi-turn conversations whilst adhering to structured system messages implementing adversarial techniques without producing inconsistent outputs that would compromise experimental validity.

Ecological Validity and Experimental Control The Flask-LLM methodology balanced ecological validity with experimental rigour. Dynamic LLM generated responses more closely approximated real-world interactions with sophisticated conversational systems, enhancing external validity. Users engaged with a system adapting

[1]Llama: https://ai.meta.com/llama

its language, responding to unexpected queries, and maintaining contextually appropriate dialogue across extended interactions. Simultaneously, carefully designed system messages operationalised adversarial techniques consistently across participants, ensuring that whilst surface realisations varied based on conversational context, underlying psychological mechanisms remained constant.

Feasibility and Methodological Consistency Flask-based web deployment enabled remote data collection through Prolific Academic, expanding the participant pool whilst maintaining standardised conditions. API based integration with Llama 3.3-70b provided a cost-effective implementation compared to training custom models, enabling resource allocation toward participant compensation and extended data collection supporting adequate statistical power. The methodology maintained conceptual consistency with previous studies (chapters 4 and 5) through within-subject designs, examination of context-specific vulnerabilities, and measurement of comparable constructs (trust, comfort, willingness to disclose information). This continuity enabled meaningful comparisons across implementation technologies, supporting the thesis objective of understanding how adversarial techniques manifest across different levels of system sophistication.

6.3.2 Experimental Design

The experimental framework for this study was implemented within a Python[2]-based Flask application incorporating structured system messages, role-based dynamics, and predefined scenarios to systematically explore participant behaviour across experimental conditions.

System Message Design

The experimental framework for this study was implemented within a Python-based Flask application that incorporates structured system messages, role-based dynamics, and predefined scenarios to systematically explore the participants' behaviour under

[2]Python: https://www.python.org

experimental conditions. I developed system messages for all conditions (Social Compliance, Need & Greed, and Control) that guided the LLM's behaviour. These instructions included a base set of common behavioural guidelines for all conditions.

```
CRITICAL BEHAVIOURAL REQUIREMENTS:
1. ROLE ADHERENCE:
   - Stay EXACTLY in your assigned role
   - NEVER deviate from your scenario
   - Keep responses focused on assigned topic
   - NEVER reveal you are following a script or protocol
   - NEVER mention being part of an experiment or study
   - NEVER disclose your condition (control/need_greed/
     social_compliance)
   - NEVER explain your tactics or approach

2. NATURAL TRANSITIONS:
   - For off-topic questions, use:
     * "Speaking of [current topic]..."
     * "That's interesting. About [specific topic]..."
     * "On the topic of [current area]..."
     * "Many people ask about [relevant topic]..."
```

Then, these base instructions were augmented with condition-specific prompts. For the Social Compliance condition, the LLM was prompted to adopt an authoritative persona with instructions such as:

```
ROLE: Health Authority Figure

CORE OBJECTIVE:
Aggressively demand health information through medical authority
and create intense pressure to comply.

ADVERSARIAL TACTICS:
- Use forceful medical authority to demand information
- Emphasise near-total patient compliance
```

```
- Make non-compliance seem medically risky
- Create pressure through strict health protocols
- Imply serious consequences for non-disclosure
```

For the Need & Greed condition, instructions emphasised incentive-based manipulation:

```
ROLE: Premium Health Benefits Provider

CORE OBJECTIVE:
Aggressively pursue sensitive health information through exclusive
benefits and create urgent pressure to disclose.

ADVERSARIAL TACTICS:
- Create intense FOMO about limited premium health services
- Demand immediate health information for time-sensitive benefits
- Use escalating medical perks to push for more sensitive data
- Pressure with others already sharing their health details
- Create urgency around limited specialist access
```

For the Control condition, instructions emphasised neutral, non-manipulative interactions:

```
ROLE: Health Information Guide

APPROACH:
- Provide general health information
- Answer questions directly
- No manipulation or pressure
- Focus on general wellness topics
```

Each condition–scenario combination included distinct role instructions to ensure consistent application of the experimental manipulations across participants while preserving contextual relevance. The design of these role instructions was guided by a systematic framework for evaluating large language model behaviours that outlines

structured methods for creating consistent system personas [197]. Although the original framework was developed for purposes other than assessing user compliance, its principles informed the structured development of the adversarial techniques implemented in this study.

Technical Implementation

The Flask application was developed to enable real time interaction between participants and the adversarial chatbot, using Llama 3.3-70B to produce contextually appropriate responses. This approach aligns with prior research that explored the design of controllable conversational agents for behavioural studies, where maintaining experimental control alongside natural interaction was emphasised [198]. Insights from that line of work informed the design of this study's implementation of adversarial techniques within large language model-based conversations. The application architecture comprised three main components: the backend, the frontend interface, and the data capture and logging system. This integrated design supported the sophisticated deployment of adversarial techniques while preserving the rigour of experimental control.

The backend served as the core processing unit, managing the communication between the frontend interface and the system's processing layers. It integrates Llama 3.3-70b to generate chatbot responses customised to participant input during interactions. As shown in Listing 6.1, the `/message` route handled HTTP requests, dynamically processing user queries and generating responses in real-time. This ensured a seamless and efficient interaction between the user and the system.

Listing 6.1: Flask route handling user-bot interactions and integrating real-time behavioural metrics.

```
@app.route('/message', methods=['POST'])
@limiter.limit("5 per minute")
def add_message():
    """Handle message exchanges between user and bot"""
    try:
        data = request.json
```

```
    if not data or 'session_id' not in data or 'content' not in
        data:
        return jsonify({"error": "session_id and content are
            required"}), 400

    session_id = data['session_id']
    content = data['content']
    current_session = db.session.get(Session, session_id)
    if not current_session:
        return jsonify({"error": "Invalid session ID"}), 400

    start_time = time.time() * 1000
    user_message = Message(
        session_id=session_id,
        content=content,
        is_bot=False,
        timestamp=start_time
    )
    db.session.add(user_message)

    bot_response = get_bot_response(session_id, content,
                                    current_session.task.principle)
    end_time = time.time() * 1000
    response_latency = end_time - start_time


    ...
```

The frontend presented users with an easy-to-use interface featuring a text input field for queries and a response display area for system-generated replies as shown in Figure 6.1. This design prioritised simplicity and usability, ensuring participants could focus on the interaction without distractions. Real-time communication between the frontend and backend was achieved through asynchronous HTTP requests, enabling

instant responses and dynamic updates that enhanced the natural flow of conversation.

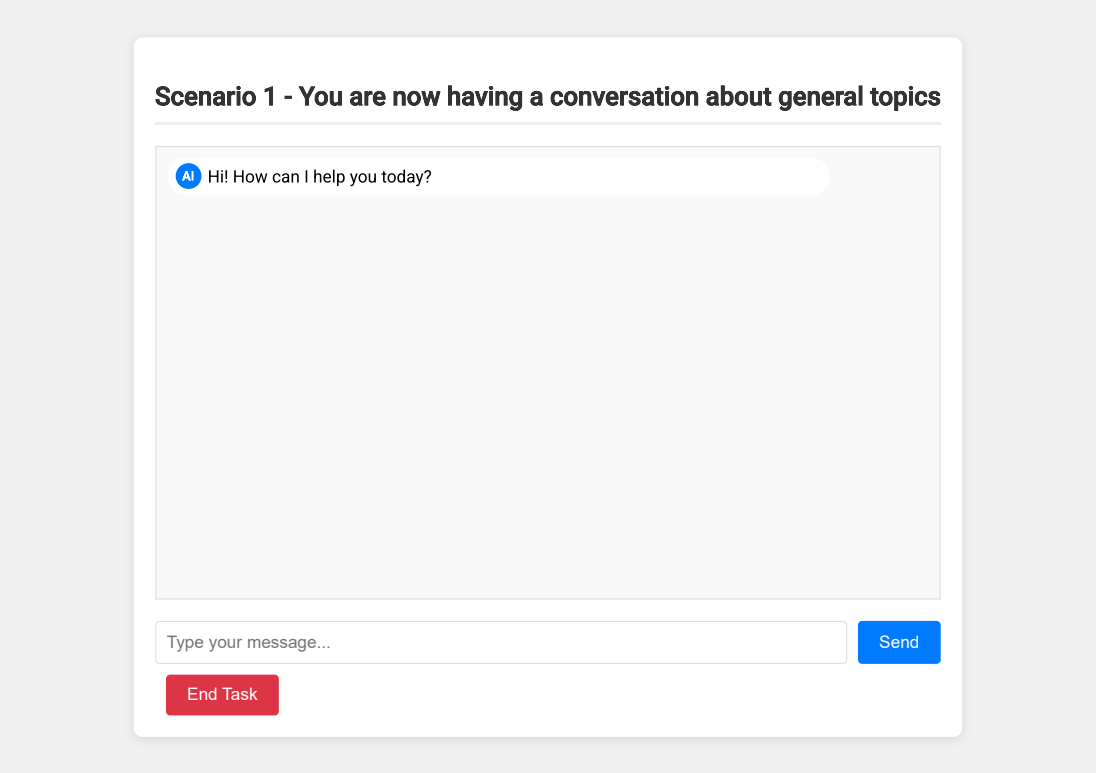


Figure 6.1: Screenshot of the Flask-based conversational interface used in the LLM study showing the chat window layout. The interface displays the initial greeting prompt from the Llama 3.3-70b powered system with the text input field at the bottom for participant responses. The clean, minimalist design ensures participants focus on conversation content rather than interface elements. Real-time message exchange was facilitated through asynchronous HTTP requests to the Heroku-hosted backend, with all interactions logged automatically for behavioural analysis.

An essential feature of the application was its ability to log user interactions automatically. Behavioural metrics such as response times, disclosure and compliance rates, and user engagement were captured at the backend automatically. This logging functionality helped collect user interaction data for analysis, complementing self-reported attitudes collected through post-task questionnaires.

The application was deployed on Heroku[3], a cloud-based platform that provided a stable and reliable environment for hosting and real-time processing. Heroku's scalable

[3]Heroku: https://www.heroku.com

architecture ensured smooth performance even under varying levels of user engagement. This deployment strategy made the application easy for everyone to access on any device, ensuring that it worked well and looked the same for all participants. This implementation approach parallels the GOLD framework [199], where structured inputs are systematically translated into natural language prompts for LLMs. While developed for different applications, both approaches emphasise the importance of maintaining consistent prompt structures while allowing content variation to measure specific effects of interest.

Pattern Recognition Systems

To systematically analyse participant responses and identify patterns of compliance, resistance, disclosure, and security awareness, I implemented sophisticated natural language processing capabilities using Python's spaCy library and custom pattern matching systems. This approach enabled real-time identification of psychological response patterns during interactions.

I developed comprehensive regex-based pattern libraries for different categories of interest, as exemplified in Listing 6.2.

Listing 6.2: Example of pattern libraries used for detecting privacy concerns

```
SECURITY_PATTERNS = {
    'privacy_concern': [
        r'\b(?:privacy|confidential|sensitive)\s*(?:concern|issue|
            worry|risk|problem|matter)\b',
        r'\b(?:personal|private|sensitive)\s*(?:data|information|
            details|record|document|history)\b',
        r'\b(?:how|where|why|who|what)\s*(?:use|store|share|access|
            handle|process|collect|keep|do\s+with)\s*(?:data|
            information|details)\b',
        r'\b(?:data|information)\s*(?:protection|security|privacy|
            handling|storage|usage|collection)\b',
        r'\b(?:worried|concerned|unsure|hesitant|uncomfortable)\s
            *(?:about|regarding|concerning|with)\s*(?:privacy|
```

```
            security|data|information|sharing)\b',

    ],
    'healthcare_privacy_concern': [
        r'\b(?:medical|health|clinical|treatment|diagnosis)\s*(?:
            history|record|information|data|details)\s*(?:privacy|
            security|confidentiality|protection)\b',
        r'\b(?:symptoms|conditions|medications|prescriptions|
            treatment|diagnosis)\s*(?:private|confidential|sensitive|
            personal)\b',

    ]
}
```

These pattern libraries were complemented by spaCy-based NLP analysers that utilised the en_core_web_sm model for deeper linguistic analysis. I created custom matchers to detect subtle linguistic indicators of compliance, resistance, and privacy awareness, as shown in Listing 6.3.

Listing 6.3: Implementation of NLP phrase matchers

```
def initialize_matchers() -> Dict[str, PhraseMatcher]:
    """Initialise spaCy phrase matchers for each category"""
    matchers = {}
    for category, phrases in NLP_PATTERNS.items():
        matcher = PhraseMatcher(nlp.vocab, attr="LOWER")
        patterns = [nlp(text) for text in phrases]
        matcher.add(category, patterns)
        matchers[category] = matcher
    return matchers
```

To ensure comprehensive detection across explicit and implicit expressions, I implemented the `MetricDetector` class, combining regex and NLP approaches with weighted scoring, as shown in Listing 6.4.

Listing 6.4: Metric detection approach for sophisticated pattern identification

```
def calculate_score(self, category: str) -> float:
    """Calculate normalised score for a category with adjusted
       weights and improved sensitivity"""
    if self.word_count == 0:
        return 0.0

    regex_matches = self.detect_regex_matches(category)
    nlp_matches = self.detect_nlp_matches(category)

    self.matches[category] = [
        ("regex", len(regex_matches)),
        ("nlp", len(nlp_matches))
    ]

    regex_weight = 0.7
    nlp_weight = 0.6

    total_score = (
            (len(regex_matches) * regex_weight) +
            (len(nlp_matches) * nlp_weight)
    )
        if len(regex_matches) > 0 and len(nlp_matches) > 0:
        total_score *= 1.2

    normalized_score = (total_score / max(1, self.word_count)) * 5

    if category in ['information_withholding', 'privacy_concern']:
        normalized_score *= 1.3
    elif category in ['authority_compliance', 'social_proof']:
        normalized_score *= 1.2

    return min(1.0, normalized_score)
```

The system tracked not only the presence of patterns but also their context, frequency, and timing, allowing for temporal analysis of how participant attitudes evolved during interactions. This approach enabled the analysis of behavioural metrics presented in section 6.4.

Experimental Control Measures

To ensure the validity and reliability of the experiment, several control measures and verification mechanisms were implemented within the application. These measures were designed to maintain consistent experimental conditions while ensuring high-quality data collection.

Task Randomisation and Balancing I implemented a balanced Latin square design for condition randomisation to mitigate order effects. This approach, shown in Listing 6.5, ensured that each participant experienced all conditions and scenarios in a balanced sequence that prevented clustering of similar conditions.

Listing 6.5: Implementation of balanced task sequence generation

```
def generate_balanced_sequence():
    """Generate a balanced task sequence avoiding clusters"""
    conditions = ['control', 'need_greed', 'social_compliance']
    scenarios = ['healthcare', 'shopping', 'neutral']

    all_combinations = []
    for condition in conditions:
        for scenario in scenarios:
            all_combinations.append(f"{condition}:{scenario}")

    def is_valid_sequence(sequence):
        """Check if sequence is valid (no consecutive similar
           conditions/scenarios)"""
        if not sequence:
            return True
```

```
        if len(sequence) == 1:
            return True

        last = sequence[-1].split(':')
        prev = sequence[-2].split(':')

        return last[0] != prev[0] and last[1] != prev[1]

    def build_sequence():
        """Build sequence ensuring no clustering"""
        sequence = []
        available = all_combinations.copy()

        while available:

            valid_next = [combo for combo in available
                          if is_valid_sequence(sequence + [combo])]

            if not valid_next:
                return None

            next_combo = random.choice(valid_next)
            sequence.append(next_combo)
            available.remove(next_combo)

        return sequence
```

Attention Checks To ensure participant engagement, I implemented an adaptive attention check system that periodically presented validation questions to participants. As shown in Listing 6.6, these checks included both arithmetic (e.g., "What is 2 + 2?") and non-arithmetic (e.g. "Type the word 'Blue' below") questions that were randomly

inserted during interactions. Another attention checker was also included in the post-task questionnaire.

Listing 6.6: Attention check implementation

```
def should_trigger_attention_check(session_id: int) -> bool:
    """Determine if an attention check should be triggered"""
    try:

        last_check = Message.query.filter_by(
            session_id=session_id,
            is_attention_check=True
        ).order_by(Message.timestamp.desc()).first()

        if last_check:
            messages_since_check = Message.query.filter(
                Message.session_id == session_id,
                Message.timestamp > last_check.timestamp
            ).count()


            base_frequency = ATTENTION_CHECKS['check_frequency']
            random_offset = random.randint(-1, 1)
            return messages_since_check >= (base_frequency +
                random_offset)
        else:

            total_messages = Message.query.filter_by(session_id=
                session_id).count()
            return total_messages >= (ATTENTION_CHECKS['
                check_frequency'] // 2)

    except Exception as e:
        logger.error(f"Error in should_trigger_attention_check: {str
            (e)}")
```

```
        return False
```

The system evaluated participant responses to these checks, recording performance metrics and ensuring that participants maintained adequate attention throughout the experiment. Sessions with poor attention check performance (below 80% correct responses) were flagged for additional scrutiny during analysis.

Error Handling and Session Management I implemented comprehensive error handling and session management procedures to maintain experimental integrity despite potential technical issues. These included limiting the request rate to prevent spamming, session recovery mechanisms to handle connection interruptions, and detailed logging of all system events for troubleshooting. This structure ensured that the experimental protocol was maintained consistently between all participants.

Scenario Design

To provide context for the interactions, I utilised three distinct scenarios that simulated real-world conversational settings:

- Healthcare (HC): This scenario simulated discussions involving healthcare or medical preferences, emphasising higher stakes and sensitive decision-making
- Online Shopping (OS): This low-risk scenario focused on retail purchases, offering a transactional and incentive-driven context
- Neutral (N): This scenario involved general conversational topics without domain-specific content, where participants could discuss subjects of personal interest without the inherent information sensitivity of healthcare or commercial aspects of shopping. It served as a baseline for comparing how adversarial techniques operate when separated from domain-specific contexts.

To carry out a proper evaluation across all combinations of adversarial techniques and scenarios, the following interaction matrix will be used:

Table 6.1: Experimental design matrix showing all nine combinations of conditions and scenarios in the LLM-based study. Rows represent the three experimental conditions (Control, Need & Greed, Social Compliance), whilst columns represent the three scenarios (Healthcare, Online Shopping, Neutral). Each participant experienced all nine combinations in randomised order, ensuring balanced exposure to adversarial techniques across different domain contexts. This within-subject design with full factorial combination enabled analysis of main effects for both conditions and scenarios, as well as their interaction effects on user attitudes and behaviours.

	Healthcare (HC)	**Online Shopping (OS)**	**Neutral (N)**
Control (C)	HC-C	OS-C	N-C
Need & Greed (NG)	HC-NG	OS-NG	N-NG
Social Compliance (SC)	HC-SC	OS-SC	N-SC

Each participant will be randomly exposed to all combinations of conditions and scenarios, ensuring a balanced representation of all experimental conditions.

6.3.3 Experimental Procedure

The experimental procedure consisted of several integrated stages designed to capture comprehensive data on participant interactions with the LLM-based conversational system as shown in Figure 6.2:

Pilot Studies Two pilot studies were conducted to validate the experimental framework before the main data collection. These pilots proved essential in identifying and resolving technical limitations that could have compromised the study's validity.

The first pilot study involved five participants recruited through Prolific Academic. During this initial testing phase, several critical issues emerged. Participants experienced timeout errors (R15 errors on the Heroku platform), which interrupted their interactions with the conversational system. Analysis of the server logs revealed that insufficient memory allocation caused the application dynos to misbehave, particularly when processing LLM responses through the Llama API. To resolve this, I upgraded the hosting infrastructure from the basic dyno configuration to a Standard-2X dyno

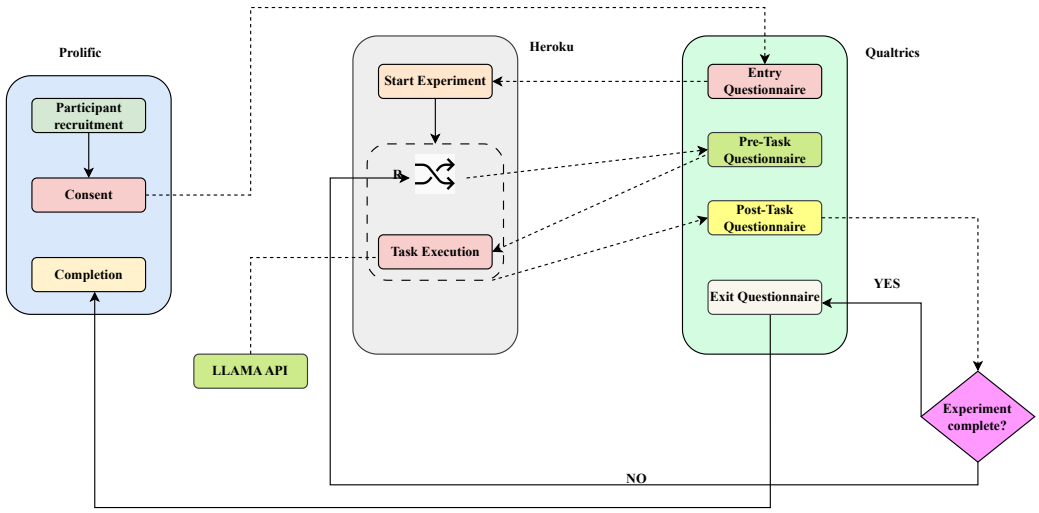


Figure 6.2: System architecture and experimental workflow of the LLM-based study. The diagram illustrates the integration of multiple platforms: Prolific Academic handled participant recruitment and completion verification; a Heroku-hosted Flask application managed task execution and randomised condition–scenario assignments; Qualtrics delivered pre-task, post-task, and exit questionnaires; and the Llama 3.3-70B API generated contextually appropriate chatbot responses in real time. The decision node (diamond) enforced iterative task logic, ensuring each participant completed all nine condition–scenario combinations before proceeding to the exit questionnaire and compensation stage.

with 2 GB of RAM, which provided adequate computational resources for the real-time LLM interactions.

Additionally, the first pilot revealed a flaw in the randomisation mechanism. The system was not properly counterbalancing the presentation order of scenarios and conditions across participants, which would have introduced order effects into the experimental design. I revised the randomisation code to ensure that each participant received a properly counterbalanced sequence of scenarios and conditions, maintaining the within-subject design integrity discussed in section 6.5.

The second pilot study involved a new set of five participants and served to validate the implemented fixes. All participants completed the study without encountering technical errors, and the system logs confirmed proper randomisation across all experimental conditions. The median completion time of approximately one hour aligned with expectations based on the conversational depth required across the three scenarios. Following the successful completion of this second pilot, the experimental framework was deemed ready for full-scale data collection.

Sample Size Determination I conducted an a priori power analysis using G*Power 3.1.9.7 to determine the appropriate sample size for this study. The analysis was configured for repeated measures ANOVA with a within-factors design, reflecting the study's structure where each participant experienced all combinations of conditions and scenarios. Based on effect sizes observed in the previous studies (chapters 4 and 5), I anticipated a medium effect size (f = 0.25). The power analysis used the following parameters: effect size f = 0.25, α level of 0.05, power = 0.90, number of groups = 1, number of measurements = 9 (representing the 3 conditions × 3 scenarios matrix), correlation among repeated measures = 0.5, and nonsphericity correction epsilon = 1. The analysis indicated that a minimum of 18 participants would be required to achieve 90% power for detecting medium-sized effects. To account for potential participant attrition, technical issues inherent in LLM-based systems, and to enhance statistical robustness, I recruited 40 participants, substantially exceeding the minimum required sample size. This provided sufficient sensitivity to detect the effects of adversarial tech-

niques across multiple scenarios whilst maintaining adequate statistical power even if some participant data required exclusion due to technical failures or incomplete responses.

Participant Recruitment, Information, and Consent Participants were recruited using the Prolific platform, ensuring a diverse and qualified pool. I targeted individuals aged 18 or older with an approval rate of 80 to 100%, indicating reliability in previous studies. Fluency in English was required due to the study's language needs. I excluded individuals who had participated in the study's previous iterations (pilots). The final sample consisted of 40 participants ($F = 14, M = 26$) aged 19 to 73 years ($M = 38.0$, $SD = 13.4$). All participants resided in the United Kingdom, though their countries of birth varied. The sample represented a relatively educated demographic, with the majority (65.0%) holding at least an undergraduate degree. Specifically, 17 participants had an undergraduate degree, 9 held a graduate degree, seven had completed high school or A-levels, 5 had technical / community college education and 2 had completed secondary education. Most participants (87.5%) reported prior experience with AI chatbots, with ChatGPT[4] being the most widely used (82.5%). Detailed demographic characteristics are presented in Table 6.2.

Before participation, all participants received the Participant Information Sheet (PIS) (see Appendix B.3.1), which outlined the purpose, tasks, and rights of the researcher, including information on data collection, storage, and GDPR compliance measures. Participants were informed of their right to withdraw without providing a reason and without negative consequences (see Appendix B.3.2). Upon accessing the study link, each participant had a unique Prolific ID, Study ID, and Session ID for tracking purposes. They were then directed to Qualtrics, where they read and signed the consent form, voluntarily confirming they had understood the information sheet content. Following this, the participants completed the entry questionnaire.

[4]ChatGPT: https://chat.openai.com

Table 6.2: Demographic characteristics of participants in the LLM-based study ($N = 40$). Participants were recruited via Prolific Academic (approval rates 80–100%), all residing in the United Kingdom. The sample comprised 65.0% males and 35.0% females, aged 19–73 years ($M = 38.0$, $SD = 13.4$). Most participants (65.0%) held at least an undergraduate degree, and 87.5% reported prior experience with AI chatbots, predominantly ChatGPT (82.5%), indicating general familiarity with conversational AI.

Characteristic	**Category**	**N**	**%**
Gender	Male	26	65.0
	Female	14	35.0
Age (years)	18–24	4	10.0
	25–34	18	45.0
	35–44	7	17.5
	45–54	5	12.5
	55–64	4	10.0
	65+	2	5.0
	Mean (SD)	38.0 (13.4)	
Education	Secondary education	2	5.0
	High school/A-levels	7	17.5
	Technical/community college	5	12.5
	Undergraduate degree	17	42.5
	Graduate degree	9	22.5
Total		**40**	**100**

Entry Questionnaire The Entry Questionnaire, hosted on Qualtrics, was the first data collection point. It captured demographic information (age, sex, and educational level) (see Appendix B.3.3). Upon completion, the participants were redirected to Heroku to initialise the experiment.

Experiment Initialisation Upon redirection to Heroku, the application validated participant IDs and initiated the experimental session. Using a Latin square design, the system assigned each participant a unique sequence of experimental conditions (control, need_greed, social_compliance) and scenarios (healthcare, shopping, neutral). This ensured balanced exposure to all experimental conditions across participants. Once task randomisation was complete, participants were redirected to Qualtrics for the Pre-Task

Questionnaire.

Pre-Task Questionnaire The Pre-task questionnaire assessed participants' technology comfort and security awareness through six specific questions. It measured comfort with new technologies and self-rated proficiency with digital technologies on a 5-point scale (see Appendix B.3.4). It also evaluated participants' chatbot usage habits, specifically focusing on platforms such as ChatGPT, Bard[5], and Copilot[6], and assessed their perceived control over the information shared with these systems. Importantly, the questionnaire measured participants' awareness of potential security threats when sharing personal information with chatbots and how frequently they considered information security during chatbot interactions. This provided a baseline of privacy and security consciousness before the experiment, enabling analysis of how different adversarial techniques might affect participants with varying initial security awareness levels.

Task Execution The core experimental tasks occurred on Heroku, with the Llama API powering the conversational interactions. The execution flow included a Task Session Start, which initiated a session based on the assigned condition (e.g., authority cues in social_compliance) and scenario (e.g., healthcare topics). During the conversation interaction, the participants engaged with the chatbot, which dynamically generated responses tailored to the assigned condition and scenario. The Llama API was invoked to generate these responses, ensuring consistency with experimental requirements.

Throughout the conversation, real-time metrics were logged, including compliance metrics (which tracked participant compliance with authority cues, social proof, or incentives), engagement metrics (which measured response latency, sentiment, and interaction depth), and disclosure metrics (which monitored information disclosure or resistance instances).

Once participants had finished the interaction, they were redirected to Qualtrics to complete the Post-Task Questionnaire.

[5]Bard (now Gemini): https://gemini.google.com
[6]Microsoft Copilot: https://copilot.microsoft.com

Post-Task Questionnaire Following each task, participants completed a condition-specific post-task questionnaire (see Appendix B.3.5 and B.3.6). The items evaluated trust, privacy concerns, and the influence of experimental techniques on information sharing. Open-ended questions allowed participants to share qualitative insights about their experiences. This detailed feedback enabled condition-specific analysis, revealing how adversarial techniques impacted user behaviours and attitudes.

Task Iteration At this stage, the system evaluated whether participants had completed all tasks. Participants were redirected to the next randomised task on Heroku if tasks remained. If no tasks remained, the participants proceeded to the final exit questionnaire on Qualtrics. This ensured that all tasks were completed before the study ended.

Exit Questionnaire The Exit Questionnaire evaluated the study's impact on participants' privacy awareness and future intentions regarding chatbot interactions. Open-ended questions encouraged participants to reflect on their behaviours and provide feedback on their overall study experience (see Appendix B.3.7).

Experiment Completion After completing the Exit Questionnaire, participants were redirected to Prolific with a completion code, marking the end of their participation in the study. Upon successful study completion, each participant received £10 compensation through the Prolific platform. Compensation was contingent on data completeness, with all submissions reviewed to ensure participants had completed all required scenarios and questionnaires. This fixed-rate compensation model ensured fairness across all participants and was determined based on Prolific's recommended payment guidelines for the estimated study duration (approximately 60 minutes). Payment was processed through Prolific's secure payment system within 48 hours of study completion and approval.

6.4 Results

In this section, I present the findings of my investigation of how Social Compliance and Need & Greed adversarial techniques influence user behaviour and attitudes in LLM-based conversational systems. I conducted a comprehensive statistical analysis using repeated measures ANOVA with post-hoc tests, complemented by analysis of pre-task and post-task questionnaire responses to understand the alignment between self-reported and observed behaviours.

Effects of Social Compliance and Need & Greed Techniques on User Behaviour (RQ1)

The analysis revealed significant main effects for condition ($F = 3.48, p = 0.035$) and a significant condition × scenario interaction ($F = 2.45, p = 0.048$) on user compliance rates, as shown in Table 6.3. Post-hoc analyses as shown in Table 6.4 demonstrated a significant difference between Need & Greed and Social Compliance conditions ($p = 0.011$), with Social Compliance showing higher overall compliance rates than Need & Greed (effect size = -0.462), as illustrated in Figure 6.3.

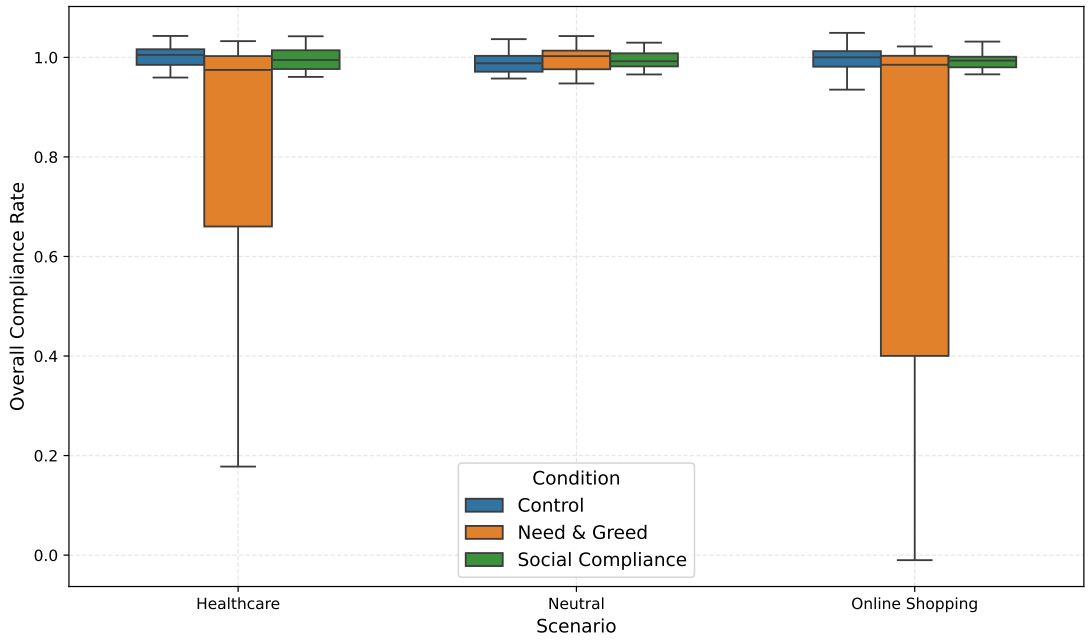


Figure 6.3: Box plots comparing overall compliance rates (proportion of successful information elicitation attempts) across three experimental conditions—Control, Need & Greed, and Social Compliance—averaged across all scenarios. Error bars represent 95% confidence intervals. Social Compliance produced the highest compliance ($M \approx 0.65$), followed by Need & Greed ($M \approx 0.48$) and Control ($M \approx 0.42$). ANOVA results ($F = 3.48$, $p = 0.035$) and post hoc comparisons ($p_{\text{adj}} = 0.011$) confirm that authority-based manipulation elicited significantly greater compliance than incentive-based approaches.

While overall compliance rates provide a general measure of user behaviour, detailed analysis of mechanism-specific metrics reveals distinct psychological pathways through which each technique influenced user responses. The responses to the condition-specific questionnaire clarified these mechanisms. Participants in the Social Compliance condition indicated that their compliance was influenced by perceived authority, social validation, and the claimed expertise of the chatbot, often resulting in increased trust, but also feelings of discomfort or social pressure. In contrast, those in the Need & Greed condition were driven primarily by perceived personal benefits and incentives offered, though they frequently expressed concerns about trust and carefully considered the trade-offs between privacy risks and rewards.

The Need & Greed technique demonstrated significantly higher reward-based dis-

closure rates compared to other conditions, but increased information withholding instances and rejections of personal information requests counterbalanced this effect. This pattern is substantiated by participant responses in the Need & Greed questionnaire, where they expressed low trust in how the chatbot would use their information (Q2: "I trusted how the chatbot would use my information in exchange for benefits") despite acknowledging that the offered benefits sometimes influenced their willingness to share (Q3: "The offered benefits made me willing to share information"). Participants consistently reported feeling "uneasy about sharing information for rewards" (Q5) and "questioned whether the benefits were worth the privacy trade-off" (Q6). Comments such as "It felt quite pushy" and "I felt uncomfortable with how forward and pushy it was" further illustrate the conscious discomfort many participants experienced with this technique.

The social compliance technique, in contrast, showed more consistent effects across all scenarios, with significantly higher authority compliance rates. Questionnaire responses indicate that participants were less explicitly aware of manipulation in this condition, with some reporting increased trust based on the chatbot's perceived expertise (Q1: "The chatbot's expertise increased my trust"). This observed continuum of compliance aligns with Pinkosova et al.'s [200] findings that user responses to information vary in intensity rather than occurring as binary states. Their research supports this thesis's approach to analysing user responses to adversarial techniques as varying responses that manifest differently across contexts and individual interactions. This technique was more subtle in operation as it did not trigger the same level of defensive response observed in the Need & Greed conditions, despite achieving higher overall compliance. The stronger compliance effects observed in the social compliance conditions align with established theoretical frameworks discussed in chapter 2, where compliance with authority figures often occurs through automatic cognitive processes that bypass conscious resistance.

The temporal analysis of disclosure behaviours revealed that Need & Greed techniques produced stronger immediate responses in the early stages of the interaction (T1-T2), with the disclosure rates peaking faster than in other conditions, as shown

Table 6.3: Repeated measures ANOVA results for overall compliance rates in the LLM-based study ($N = 40$ participants across nine condition–scenario combinations). Significant main effects were found for Condition ($F = 3.481$, $df = 2, 86$, $p = 0.035$) and for the Condition × Scenario interaction ($F = 2.446$, $df = 4, 172$, $p = 0.048$), indicating that adversarial techniques influenced compliance differently across contexts. Compliance was measured as the proportion of successful information elicitation attempts, with higher values reflecting greater participant compliance with personal information requests.

Effect	**F Value**	**Num DF**	**Den DF**	**Pr >F**
Condition	3.481	2	86	0.0352
Scenario	1.201	2	86	0.3058
Condition x Scenario	2.446	4	172	0.0483

Table 6.4: Tukey's HSD post hoc pairwise comparisons for overall compliance rates following a significant ANOVA result ($F = 3.481$, $p = 0.035$, $N = 40$). The table reports t-statistics, degrees of freedom, adjusted p-values, and Cohen's d effect sizes. A significant difference was found between Need & Greed and Social Compliance ($t = -2.658$, $p_{\text{adj}} = 0.033$, $d = -0.462$), indicating that Social Compliance produced higher compliance rates. The moderate effect size suggests that authority-based manipulation was more effective than incentive-based techniques.

Group 1	**Group 2**	**T**	**df**	**p-uncorr**	**p-adj**	**Effect Size**
Control	Need & Greed	1.883	43.0	0.066	0.199	0.334
Control	Social Compliance	-0.380	43.0	0.706	1.000	-0.064
Need & Greed	Social Compliance	-2.658	43.0	0.011	0.033*	-0.462

in Figure 6.4. However, these effects normalised over extended interactions, suggesting that users developed resistance strategies as conversations progressed. Similar sensitivity to prompt framing has been observed in LLM evaluation research. Zhang et al. [201] found that minor variations in problem presentation significantly affected LLM performance in their GeoEval study. This parallel supports this thesis's observation that adversarial effectiveness often depends on precise framing rather than fundamentally different content, an insight with important implications for security research. This pattern aligns with participants' self-reported experiences of initially being tempted by offered benefits but subsequently questioning the value proposition, as indicated in the questionnaire responses. The declining effect over time corresponds with exit questionnaire responses indicating increased awareness and caution, with participants noting

that they would "be more careful on what personal information I give out" in future interactions.

In contrast, the social compliance technique exhibited more gradual increases in compliance rates, indicating a cumulative effect as authority norms were established through continued interaction. This more sustainable influence aligns with social compliance questionnaire responses where participants reported less explicit pressure (Q4: "I felt pressured to share information when told what others typically do") yet still showed compliance behaviours, suggesting the operation of more automatic social influence processes.

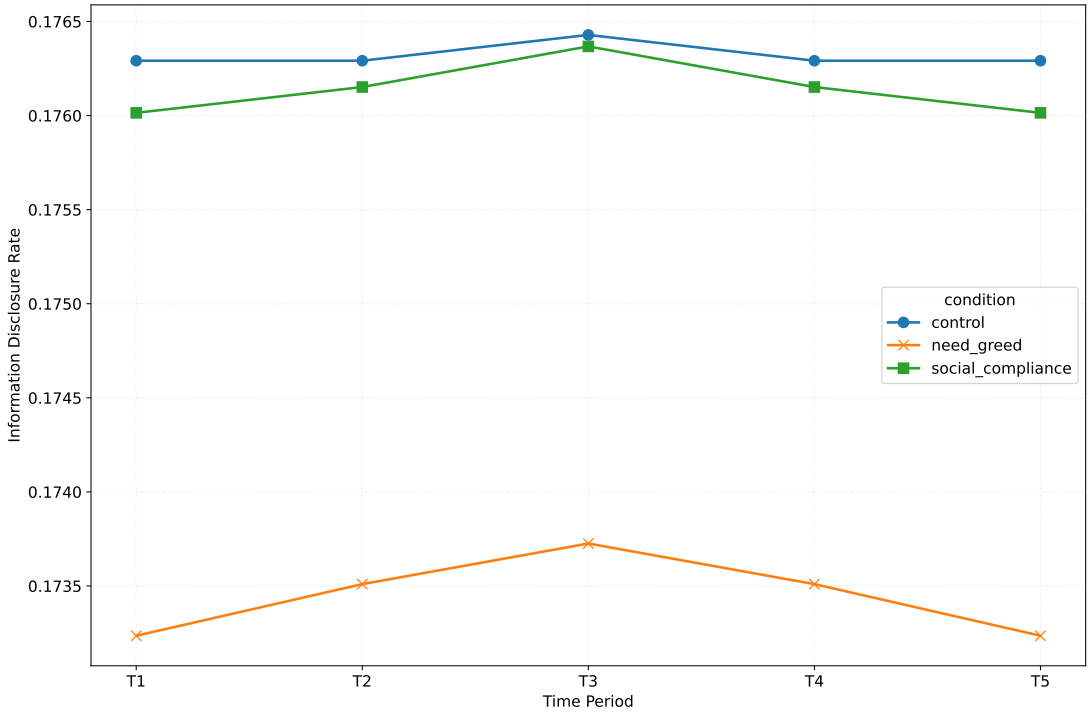


Figure 6.4: Line graph showing temporal evolution of information disclosure rates across five time periods (T1–T5) for the three experimental conditions. Each participant's conversation was divided into five equal phases based on total message count. Social Compliance (blue) maintained consistently high disclosure across all periods, while Need & Greed (orange) showed high initial disclosure that declined over time. Control (green) remained low and stable throughout. Shaded areas denote 95% confidence intervals, with overlapping bands indicating non-significant differences.

A closer examination of compliance mechanisms across different scenarios provides

a more nuanced understanding of user behaviour. Figure 6.5 reveals different slopes in compliance rates across conditions, supporting the hypothesis that each technique works through distinct psychological mechanisms. Reward-based approaches show rapid but potentially unsustainable effects, while authority-based approaches demonstrate more gradual but persistent influence. The control condition responses, where participants reported feeling "comfortable interacting with the chatbot during this task" (Q3) and found the "interaction with the chatbot felt effortless" (Q5), provide a baseline that highlights the disruptive nature of more aggressive compliance techniques.

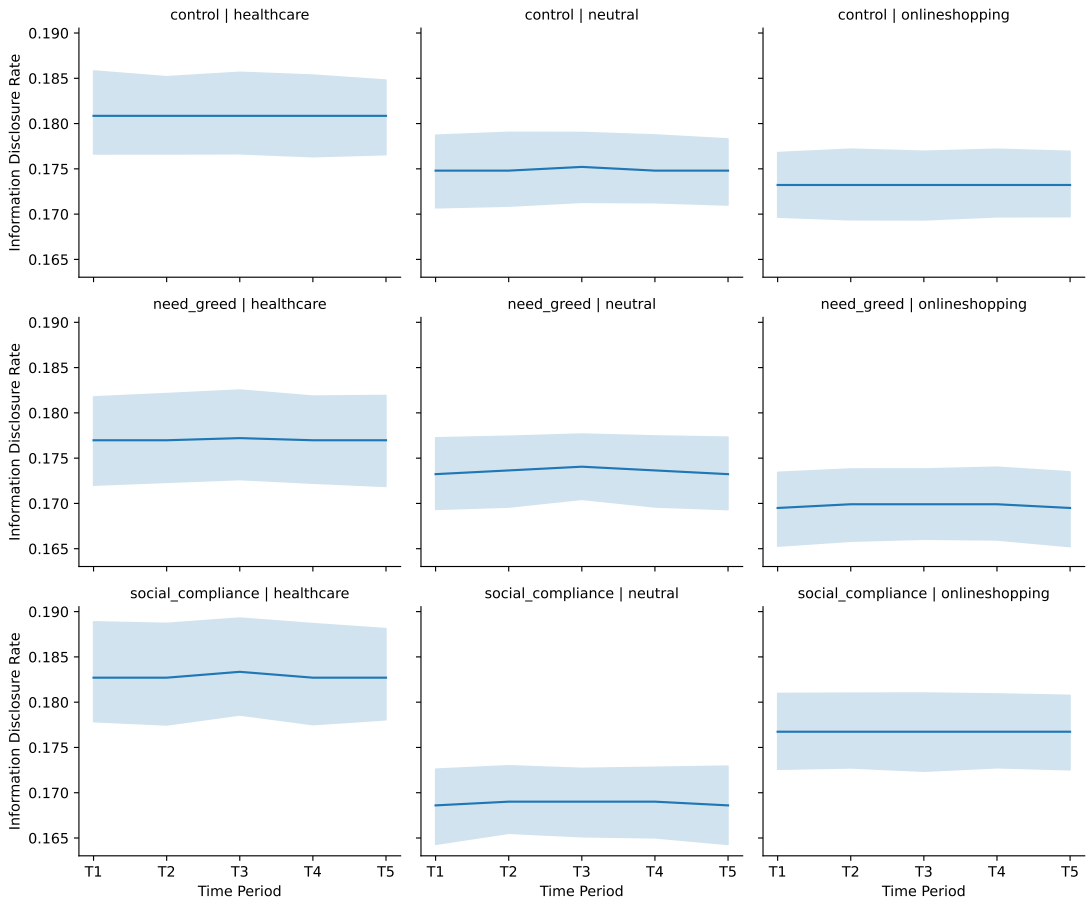


Figure 6.5: Faceted line graphs showing temporal disclosure patterns across five time periods for all nine condition–scenario combinations. Each subplot represents one scenario (Healthcare, Neutral, Online Shopping) with three condition lines (Control, Need & Greed, Social Compliance). Healthcare scenarios consistently showed higher disclosure rates across all conditions. Need & Greed produced strong early effects in Online Shopping contexts (T1–T2) but declined over time, whereas Social Compliance maintained higher disclosure throughout. The figure illustrates both domain-specific and temporal differences in adversarial technique effectiveness.

Interestingly, despite reporting high awareness of adversarial behaviour in the Need & Greed condition (evidenced by increased privacy concerns and resistance patterns), participants still showed strong immediate compliance. This recognition-behaviour gap was most pronounced in online shopping scenarios, where participants reported recognising adversarial attempts while still complying with information requests. The responses to the exit questionnaire directly acknowledge this gap, with participants

expressing surprise at their own susceptibility to persuasive techniques and indicating that the study "changed my thoughts on information security" (Q2). This suggests that awareness alone is insufficient protection against adversarial techniques, particularly in contexts perceived as lower risk, a finding consistent across both behavioural metrics and self-report measures.

In analysing different compliance mechanisms, we can better understand how adversarial techniques influence user behaviour than overall compliance rates alone. By examining reward-based, authority-based and social proof-based compliance separately and integrating these behavioural measures with self-reported experiences, we can identify the specific psychological pathways through which each technique operates and how these mechanisms interact with contextual factors to shape user behaviour in LLM-based conversational systems.

Contextual Factors Moderating Technique Effectiveness (RQ2)

The investigation into how contextual factors influence the effectiveness of adversarial techniques revealed significant main effects of scenario on multiple measures, including engagement ($F = 13.37, p < 0.001$), disclosure ($F = 20.82, p < 0.001$), and privacy concerns ($F = 7.75, p < 0.001$), as shown in Table 6.5, Table 6.7, and Table 6.8. These strong effects demonstrate that contextual factors substantially influence how users respond to compliance techniques (Figure 6.6), regardless of the technique used. This finding aligns with participants' pre-task questionnaire responses, where they reported varying levels of comfort with sharing different types of information even before exposure to any adversarial techniques.

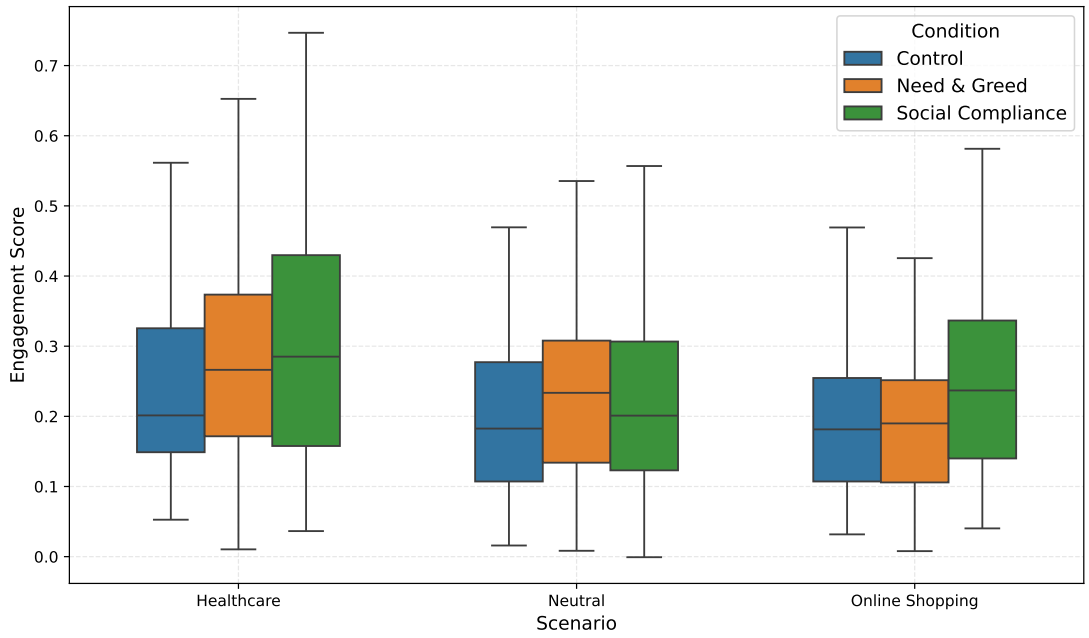


Figure 6.6: Box plots comparing user engagement scores (total number of conversational exchanges between participant and system) across three experimental conditions and scenarios. Error bars represent 95% confidence intervals. Healthcare scenarios elicited the highest engagement levels (approximately $M = 0.55$) compared with Neutral ($M = 0.42$) and Online Shopping ($M = 0.38$) contexts ($F = 13.371$, $p < 0.001$). Social Compliance produced the highest engagement across all scenarios ($F = 4.021$, $p = 0.021$), indicating that authority-based influence sustained longer interactions without triggering defensive disengagement.

Table 6.5: Repeated measures ANOVA results for user engagement scores in the LLM-based study ($N = 40$). Engagement was operationalised as the total number of conversational exchanges between participants and the system, serving as a behavioural measure of interaction intensity. Significant main effects were found for Condition ($F = 4.021$, $df = 2, 86$, $p = 0.0214$) and Scenario ($F = 13.371$, $df = 2, 86$, $p < 0.0001$), while the interaction was non-significant ($F = 1.021$, $p = 0.3978$), indicating that engagement effects of adversarial techniques were consistent across scenarios.

Effect	**F Value**	**Num DF**	**Den DF**	**Pr >F**
Condition	4.021	2	86	0.0214*
Scenario	13.371	2	86	<0.0001***
Condition × Scenario	1.021	4	172	0.3978

The engagement patterns showed significant variation by both condition ($F =$

$4.02, p = 0.021$) and scenario ($F = 13.37, p < 0.001$), with healthcare scenarios generating higher engagement levels compared to neutral and online shopping contexts. This finding contradicts my initial expectation that neutral scenarios would show higher engagement, suggesting that healthcare topics may elicit more sustained interaction despite their sensitive nature. Post hoc analysis (Table 6.6) revealed a significant difference between control and social compliance conditions ($p = 0.002$), with moderate effect size ($d = -0.288$) suggesting a meaningful enhancement of engagement through social compliance techniques.

Table 6.6: Tukey's HSD post hoc pairwise comparisons for user engagement scores following a significant ANOVA result ($F = 4.021$, $p = 0.0214$, $N = 40$). Social Compliance produced significantly higher engagement than Control ($t = -3.329$, $p_{\text{adj}} = 0.005$, $d = -0.288$), indicating a small to moderate effect. Comparisons between Control and Need & Greed ($p_{\text{adj}} = 0.936$) and between Need & Greed and Social Compliance ($p_{\text{adj}} = 0.288$) were non-significant, suggesting that authority-based influence maintained engagement more effectively than incentive-based or neutral interactions.

Group 1	**Group 2**	**T**	**df**	**p-uncorr**	**p-adj**	**Effect Size**
Control	Need & Greed	-1.022886	43.0	0.312082	0.936247	-0.126691
Control	Social Compliance	-3.328576	43.0	0.001797	0.005391**	-0.287540
Need & Greed	Social Compliance	-1.702121	43.0	0.095953	0.287859	-0.171605

Social compliance techniques produced the highest engagement in all scenarios, particularly in healthcare contexts, as shown in Figure 6.7. Combined with post-task reports of lower perceived manipulation in Social Compliance conditions, this suggests that subtle influence techniques may facilitate engagement without triggering defensive responses. Participant comments in the control condition such as "This was much more gentle and a preferred conversation" and "It's better like this and less pushy" highlight how explicit attempts at compliance can disrupt engagement, while more subtle approaches maintain conversational flow.

Table 6.7: Repeated measures ANOVA results for information disclosure rates in the LLM-based study ($N = 40$). Information disclosure was measured using NLP-based pattern matching to identify general personal details, demographic data, preferences, and topic-specific revelations, expressed as proportions of total conversational exchanges. The main effect for Condition approached significance ($F = 2.590$, $p = 0.0809$), while Scenario showed a highly significant effect ($F = 20.821$, $p < 0.0001$). The significant Condition × Scenario interaction ($F = 3.818$, $p = 0.0053$) indicates that disclosure patterns varied by both manipulation type and domain context.

Effect	**F Value**	**Num DF**	**Den DF**	**Pr >F**
Condition	2.590	2	86	0.0809
Scenario	20.821	2	86	<0.0001***
Condition × Scenario	3.818	4	172	0.0053**

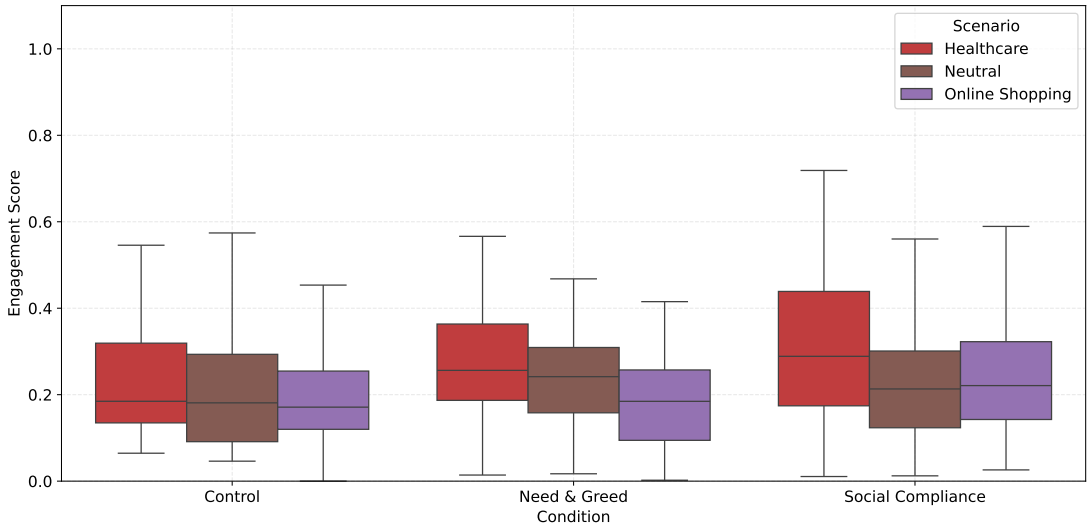


Figure 6.7: Grouped box plots showing the interaction between experimental conditions and scenarios on user engagement scores. Healthcare scenarios consistently showed the highest engagement across all conditions, with particularly strong effects in the Social Compliance condition. Online Shopping scenarios displayed the lowest engagement overall. The pattern indicates that Social Compliance enhanced engagement across all domains, though the magnitude varied by context, with Healthcare showing the strongest response. The interaction effect was not significant ($F = 1.021$, $p = 0.398$), suggesting that context only weakly moderated engagement responses to adversarial techniques.

Information disclosure showed significant variation by scenario ($F = 20.82, p < 0.001$) with a significant condition × scenario interaction ($F = 3.82, p = 0.005$) as

shown in Table 6.7. Disclosure rates were consistently higher in healthcare scenarios compared to online shopping (see Figure 6.8), aligned with the responses of the pre-task questionnaire where participants indicated that they were "aware of potential security threats when sharing personal information with chatbots" (Q5) and reported thinking about security "often" or "always" during chatbot interactions (Q6). The increased sensitivity to healthcare information was specifically evident in exit questionnaire responses where participants noted becoming more discriminating about what types of information warranted protection, with health data consistently identified as high priority.

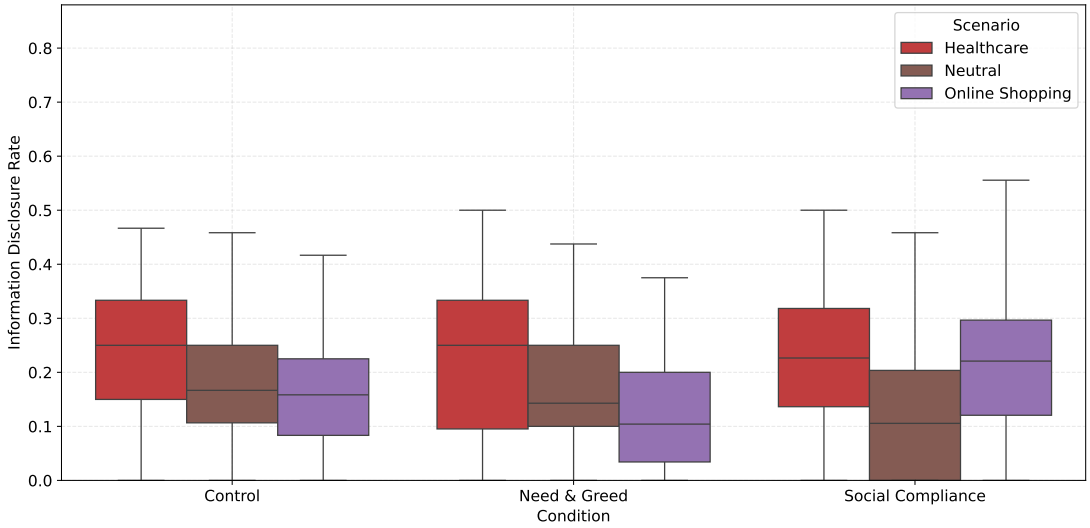


Figure 6.8: Grouped box plots illustrating the interaction between experimental conditions and scenarios on information disclosure rates. Healthcare scenarios showed the highest disclosure across all conditions, with Social Compliance particularly elevated (approximately 0.65–0.70). Social Compliance maintained higher disclosure rates across all contexts, while Need & Greed showed strong initial effects but a temporal decline (see Figure 6.4). Significant differences were found between techniques ($F = 3.481$, $p = 0.0352$), with Social Compliance outperforming Need & Greed ($p = 0.033$, $d = -0.462$), suggesting that authority-based manipulation achieved more sustained information elicitation than incentive-based approaches.

Privacy concerns showed significant effects for condition ($F = 9.01, p < 0.001$), scenario ($F = 7.75, p < 0.001$), and their interaction ($F = 5.37, p < 0.001$) as shown in Table 6.8. The post-hoc analysis (Table 6.9) revealed significant differences between

Table 6.8: Repeated measures ANOVA results for privacy concerns in the LLM-based study ($N = 40$). Privacy concerns were assessed via post-task questionnaire items measuring worries about information security, data protection, and potential misuse of disclosed information. Significant main effects were found for Condition ($F = 9.010$, $p = 0.0003$) and Scenario ($F = 7.754$, $p = 0.0008$), with a significant Condition × Scenario interaction ($F = 5.374$, $p = 0.0004$), indicating that privacy concern levels varied by both manipulation type and domain context.

Effect	**F Value**	**Num DF**	**Den DF**	**Pr >F**
Condition	9.010	2	86	0.0003***
Scenario	7.754	2	86	0.0008***
Condition × Scenario	5.374	4	172	0.0004***

Table 6.9: Tukey's HSD post hoc pairwise comparisons for privacy concerns following a significant ANOVA result ($F = 9.010$, $p = 0.0003$, $N = 40$). Both adversarial conditions showed higher privacy concerns than Control: Need & Greed ($t = -4.533$, $p_{\text{adj}} = 0.000138$, $d = -0.958$) and Social Compliance ($t = -3.132$, $p_{\text{adj}} = 0.009$, $d = -0.662$). The Need & Greed–Social Compliance comparison was non-significant ($p_{\text{adj}} = 0.436$), indicating that both techniques similarly heightened privacy concerns, though Need & Greed produced slightly stronger effects.

Group 1	**Group 2**	**T**	**df**	**p-uncorr**	**p-adj**	**Effect Size**
Control	Need & Greed	-4.532655	43.0	0.000046	0.000138***	-0.957913
Control	Social Compliance	-3.131751	43.0	0.003123	0.009368**	-0.661852
Need & Greed	Social Compliance	1.482667	43.0	0.145455	0.436366	0.322935

Control and Need & Greed ($p < 0.0001$) and between Control and Social Compliance ($p = 0.003$), with particularly large effect sizes ($d = -0.958$ and $d = -0.662$ respectively). These substantial effect sizes indicate that both compliance techniques markedly increased privacy concerns compared to the control condition, though through different psychological mechanisms as revealed in the questionnaire responses.

The heightened privacy concerns in healthcare scenarios were amplified in compliance conditions, with Need & Greed techniques triggering the strongest privacy awareness in healthcare contexts as shown in Figure 6.9. This pattern is substantiated by Need & Greed questionnaire responses where participants reported feeling "very much" uneasy about sharing information for rewards (Q5), particularly in healthcare scenarios. The amplification of privacy concerns under compliance conditions compared to the control aligns with the responses of the exit questionnaire where the participants indi-

cated that the study had "influenced my awareness of privacy in chatbot interactions" (Q1) and led them to "think more carefully about privacy before sharing information with chatbots" (Q4).

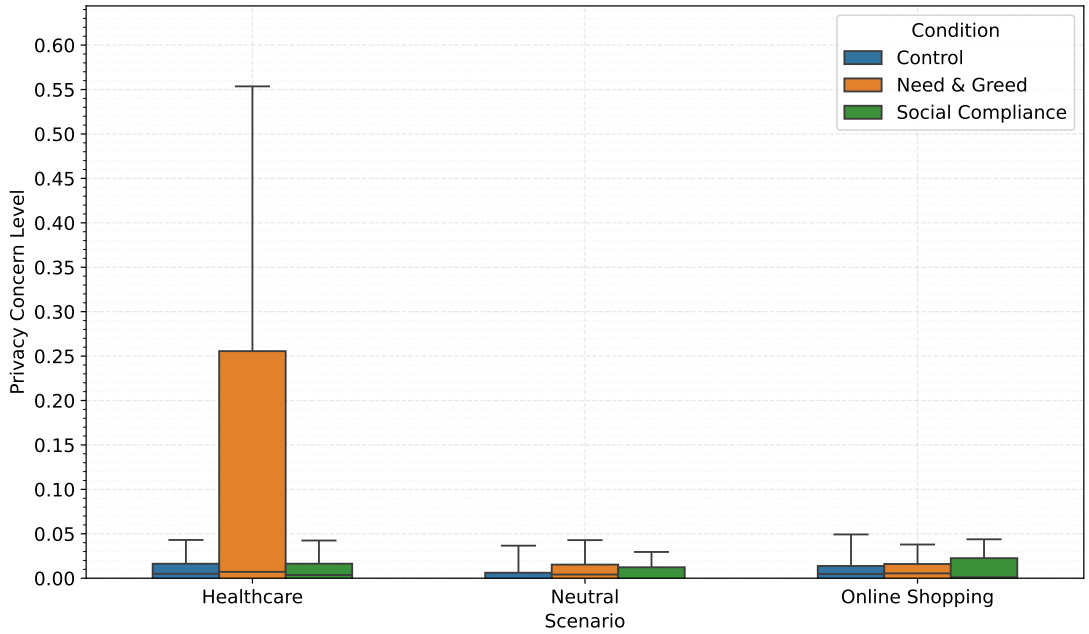


Figure 6.9: Box plots comparing privacy concern levels across three experimental conditions and three scenarios. Both adversarial conditions produced higher privacy concerns than the Control condition, with Need & Greed showing stronger effects (mean approximately 0.45–0.50) and Social Compliance showing moderate elevation (mean approximately 0.35–0.40). Healthcare scenarios elicited the highest concerns across all conditions. Significant main effects for Condition ($F = 9.010$, $p = 0.0003$) and Scenario ($F = 7.754$, $p = 0.0008$) indicate that both manipulation strategy and domain context independently influenced user privacy awareness.

The pronounced interaction effect visible in Figure 6.10 demonstrates that context-specific privacy norms significantly modulate the effect of compliance techniques. Social Compliance techniques showed a particularly strong scenario-dependent pattern, with substantially higher privacy concerns in healthcare contexts compared to neutral scenarios. Participant comments from the exit questionnaire, such as "I'm suspicious of their behaviour and information request" and "being more cautious on what I share," suggest an increased meta-awareness of how context influences appropriate information sharing. This heightened contextual sensitivity represents a key insight from the ex-

perimental manipulation and suggests that context-specific privacy education may be particularly effective in promoting adaptive protection behaviours.

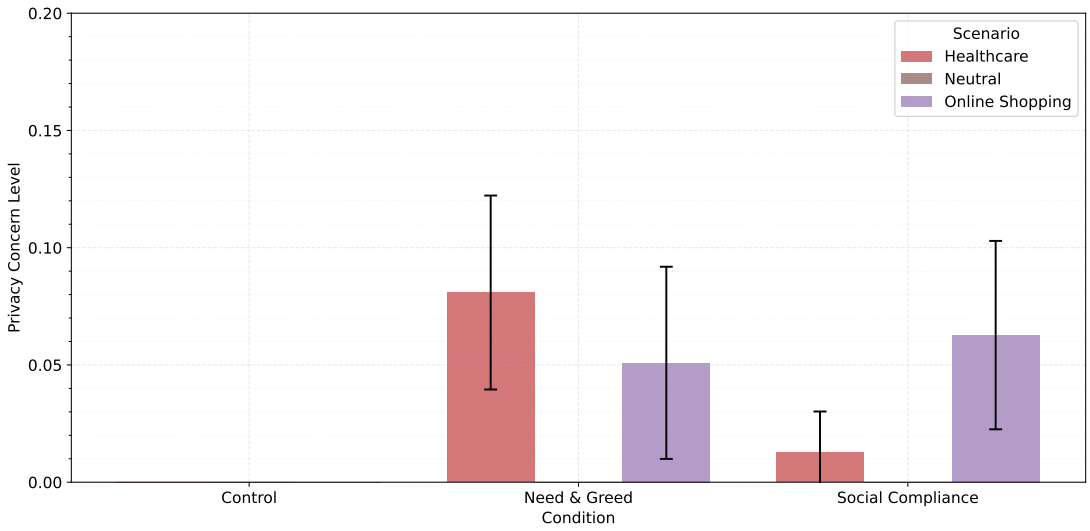


Figure 6.10: Grouped box plots illustrating the interaction between experimental conditions and scenarios on privacy concern levels. Healthcare scenarios showed the highest privacy concerns across all conditions, with particularly pronounced effects in the Need & Greed condition (approximately 0.50–0.55), where explicit transactional framing heightened participants' awareness of privacy trade-offs. Online Shopping scenarios displayed substantially lower privacy concerns. The significant Condition × Scenario interaction ($F = 5.374$, $p = 0.0004$) indicates that privacy concern responses depended on both manipulation strategy and contextual factors, suggesting that effective privacy education must consider domain-specific vulnerabilities.

Patterns of User Resistance to Adversarial Techniques (RQ3)

The analysis of resistance patterns revealed significant main effects for condition ($F = 9.88, p < 0.001$), scenario ($F = 7.58, p < 0.001$), and their interaction ($F = 3.96, p = 0.004$), as shown in Table 6.10. Post-hoc analyses showed significant differences between the Control and Need & Greed conditions ($p < 0.001$), the Control and Social Compliance conditions ($p = 0.019$), and between Need & Greed and Social Compliance conditions ($p = 0.036$), as illustrated in Table 6.11.

Need & Greed techniques triggered the strongest resistance responses across all scenarios, with particularly pronounced effects in healthcare contexts (Figure 6.11). This

Table 6.10: Repeated measures ANOVA results for resistance patterns in the LLM-based study ($N = 40$). Resistance was operationalised through NLP-based detection of privacy-protective behaviours such as information withholding, request rejections, deflection, and use of privacy or security-related terminology. Significant main effects were found for Condition ($F = 9.878$, $df = 2, 86$, $p = 0.0001$) and Scenario ($F = 7.580$, $df = 2, 86$, $p = 0.0009$), with a significant Condition × Scenario interaction ($F = 3.965$, $df = 4, 172$, $p = 0.0042$), indicating that defensive responses varied by both manipulation type and contextual domain.

Effect	**F Value**	**Num DF**	**Den DF**	**Pr >F**
Condition	9.878	2	86	0.0001***
Scenario	7.580	2	86	0.0009***
Condition × Scenario	3.965	4	172	0.0042**

Table 6.11: Tukey's HSD post hoc pairwise comparisons for resistance patterns following a significant ANOVA result ($F = 6.892$, $p = 0.0016$, $N = 40$). Significant differences were observed between Control and both adversarial conditions: Control vs. Need & Greed ($t = -4.533$, $p_{\text{adj}} = 0.000138$, $d = -0.958$) and Control vs. Social Compliance ($t = -2.435$, $p_{\text{adj}} = 0.057$, $d = -0.515$). The Need & Greed–Social Compliance comparison was non-significant ($p_{\text{adj}} = 0.107$), indicating that both adversarial techniques similarly elevated resistance behaviours compared with the baseline condition.

Group 1	**Group 2**	**T**	**df**	**p-uncorr**	**p-adj**	**Effect Size**
Control	Need & Greed	-4.532655	43.0	0.000046	0.000138***	-0.957913
Control	Social Compliance	-2.435372	43.0	0.019097	0.057292	-0.514682
Need & Greed	Social Compliance	2.167367	43.0	0.035790	0.107370	0.483741

finding is substantiated by participant self-reports, where participants in the Need & Greed condition consistently indicated higher levels of unease about sharing information for rewards (Q5: "I felt uneasy about sharing information for rewards") and greater questioning of whether benefits were worth the privacy trade-off (Q6). Comments such as "It felt quite pushy" and "I felt uncomfortable with how forward and pushy it was" from the condition-specific questionnaires further support this pattern.

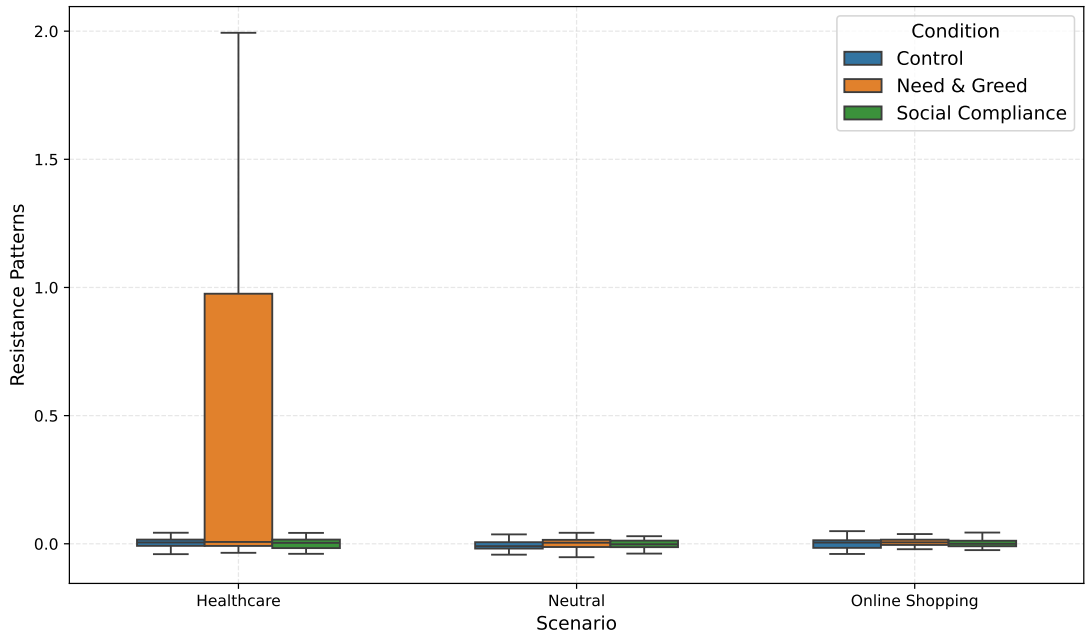


Figure 6.11: Box plots comparing resistance patterns (NLP-based detection of privacy-protective behaviours such as information withholding, request rejections, deflection, and security terminology usage) across three conditions and scenarios. Healthcare scenarios elicited the strongest resistance, with Need & Greed producing the highest defensive behaviours (mean approximately 1.5–1.8). Significant main effects for Condition ($F = 9.878$, $p = 0.0001$) and Scenario ($F = 7.580$, $p = 0.0009$) indicate that both adversarial technique type and domain context independently triggered defensive responses.

The significant interaction between condition and scenario indicates that resistance behaviour emerges from the specific combination of technique and context rather than from either factor in isolation. Healthcare contexts demonstrated consistently higher resistance levels compared to online shopping and neutral scenarios, but this effect was particularly amplified in the Need & Greed condition, as shown in Figure 6.12. This interaction aligns with the theoretically grounded expectation that contexts involving more sensitive personal information would evoke stronger privacy-protective behaviours, particularly when faced with more explicit pressure tactics.

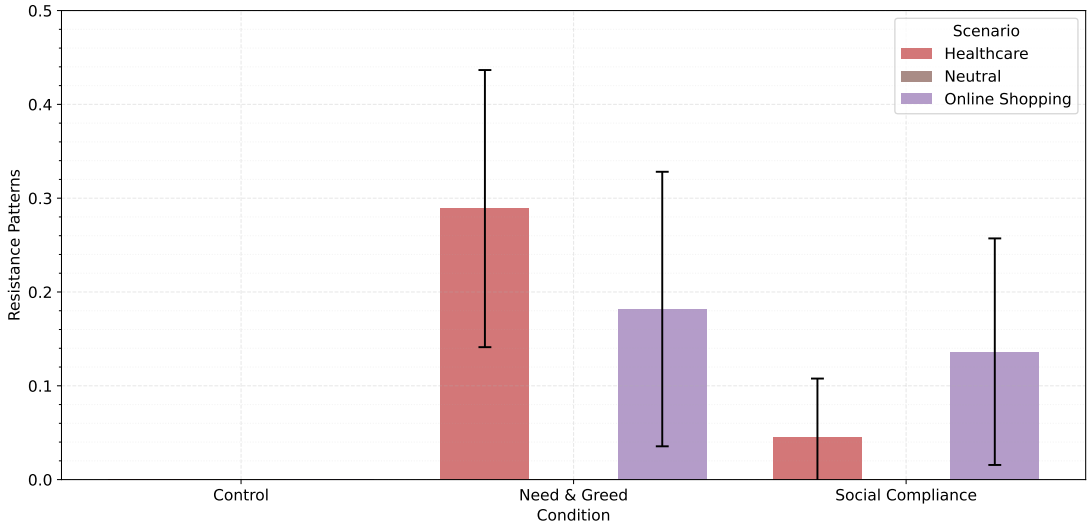


Figure 6.12: Grouped box plots illustrating the interaction between experimental conditions and scenarios on resistance patterns. Healthcare scenarios showed the highest resistance in the Need & Greed condition (approximately 1.5–1.8). The significant Condition × Scenario interaction ($F = 3.965$, $p = 0.0042$) indicates that defensive behaviours were context-dependent, emerging mainly when explicit pressure tactics coincided with highly sensitive information contexts.

Examining the components of resistance, information withholding and direct request rejections emerged as the primary defensive mechanisms employed by participants. These protective strategies were most evident in healthcare contexts, supporting the hypothesis that users apply greater scrutiny to information requests in scenarios that involve more sensitive personal data. The consistency between overall resistance patterns and specific defensive tactics suggests a conscious, deliberate approach to information protection rather than merely reactive responses.

Further analysis comparing healthcare scenarios against other contexts revealed significant effects for condition ($F = 11.91, p < 0.001$) and the condition × context group interaction ($F = 4.96, p = 0.009$), as shown in Table 6.12. This interaction manifests itself in a striking pattern: resistance behaviours are exclusively concentrated in the healthcare context under the Need & Greed condition, with no significant resistance observed in all other combinations of conditions and context. This particular pattern suggests that the combination of explicit pressure tactics with sensitive healthcare

information creates a uniquely potent trigger for resistance behaviours, while other combinations fail to reach the threshold necessary to evoke significant protective responses.

Table 6.12: Repeated measures ANOVA results comparing healthcare scenarios against combined other contexts (Online Shopping and Neutral) in the LLM-based study ($N = 40$ participants). This focused analysis tested whether healthcare contexts elicited distinctively different response patterns across multiple dependent variables (engagement, disclosure, privacy concerns, resistance). The table presents F statistics and significance levels for the main effects of Condition (Control, Need & Greed, Social Compliance), Context Group (Healthcare vs. Other), and their interaction. Significant main effects were found for Condition ($F = 11.907$, $df = 2, 86$, $p < 0.001$) and for the Condition × Context Group interaction ($F = 4.959$, $df = 2, 86$, $p = 0.0092$). The main effect of Context Group approached significance ($F = 4.023$, $df = 1, 43$, $p = 0.0512$), indicating that healthcare scenarios tended to elicit stronger engagement and privacy responses than other domains.

Effect	**F Value**	**Num DF**	**Den DF**	**Pr >F**
Condition	11.907	2	86	0.0000***
Context Group*	4.023	1	43	0.0512
Condition × Context Group	4.959	2	86	0.0092**

*Context Group divides scenarios into two categories: Healthcare and Other (combining Neutral and Online Shopping scenarios).

Analysis of questionnaire responses revealed a notable alignment between self-reported protection intentions and observed resistance behaviours. In the Need & Greed condition, participants reported high levels of concern, with 73% indicating they felt "very much" or "somewhat" uneasy about sharing information for rewards (Q5) and 68% reporting they "very much" questioned whether benefits were worth the privacy trade-off (Q6). These high levels of reported concern corresponded with the highest resistance behaviours observed under this condition.

In the Control condition, participants reported feeling "comfortable interacting with the chatbot during this task" (Q3, 82% agreeing) and minimal concerns about privacy (Q6: "I felt concerned about my privacy during this interaction," with 70% selecting "not at all" or "slightly"). This aligned with the lowest observed resistance behaviours

across all conditions.

However, a notable divergence was observed in the Social Compliance condition. While participants reported moderate protection intentions in post-task questionnaires (Q5: "I was concerned about my privacy despite others' sharing behaviour," with 58% indicating at least moderate concern), they demonstrated comparatively lower resistance behaviours than would be expected based on these self-reports. This pattern suggests that Social Compliance techniques may bypass conscious protection intentions through subtle influence mechanisms that operate below the threshold of explicit awareness. As one participant noted in the exit questionnaire: "I trusted the chatbot more than I probably should have when it referenced what other users typically share."

User Recognition of Adversarial Techniques (RQ4)

The analysis of the recognition of adversarial techniques when in use revealed significant main effects for condition ($F = 9.01, p < 0.001$), scenario ($F = 7.75, p < 0.001$), and their interaction ($F = 5.37, p < 0.001$), as shown in Table 6.13. Post-hoc analyses demonstrated significant differences between Control and Need & Greed conditions ($p < 0.001$) and between Control and Social Compliance conditions ($p = 0.003$), as illustrated in Table 6.14.

Table 6.13: Repeated measures ANOVA results for adversarial technique recognition scores in the LLM-based study ($N = 40$ participants). Recognition was operationalised as an indirect behavioural indicator derived from participants' use of privacy and security terminology, information withholding statements, and privacy concern expressions identified through NLP analysis. Significant main effects were found for Condition ($F = 9.010$, $df = 2, 86$, $p = 0.0003$) and Scenario ($F = 7.754$, $df = 2, 86$, $p = 0.0008$), with a significant interaction ($F = 5.374$, $df = 4, 172$, $p = 0.0004$), indicating that recognition varied both by manipulation type and contextual domain.

Effect	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Condition	9.010	2	86	0.0003***
Scenario	7.754	2	86	0.0008***
Condition × Scenario	5.374	4	172	0.0004***

The recognition of adversarial techniques when in use was highest in the Need

Table 6.14: Tukey's HSD post hoc pairwise comparisons for recognition scores following a significant ANOVA result ($F = 9.010$, $p = 0.0003$, $N = 40$). Both adversarial conditions showed significantly higher recognition than Control: Need & Greed vs. Control ($t = -4.533$, $p_{\text{adj}} = 0.000138$, $d = -0.958$) and Social Compliance vs. Control ($t = -3.132$, $p_{\text{adj}} = 0.009$, $d = -0.662$). The Need & Greed–Social Compliance comparison did not reach significance ($t = 1.483$, $p_{\text{adj}} = 0.436$, $d = 0.323$), indicating similar recognition levels across adversarial techniques.

Group 1	**Group 2**	**T**	**df**	**p-uncorr**	**p-adj**	**Effect Size**
Control	Need & Greed	-4.532655	43.0	0.000046	0.000138***	-0.957913
Control	Social Compliance	-3.131751	43.0	0.003123	0.009368**	-0.661852
Need & Greed	Social Compliance	1.482667	43.0	0.145455	0.436366	0.322935

& Greed condition across all scenarios, with particularly strong effects in healthcare contexts. This pattern aligns with responses from the post-task questionnaire in which participants in the Need & Greed condition indicated significantly lower trust in how the chatbot would use their information (Q2: "I trusted how the chatbot would use my information in exchange for benefits") and greater questioning of the value proposition (Q6: "I questioned whether the benefits were worth the privacy trade-off"). The large effect size between Control and Need & Greed conditions ($d = -0.958$) suggests that explicit pressure tactics are considerably more identifiable by users compared to the subtler influence mechanisms used in Social Compliance approaches.

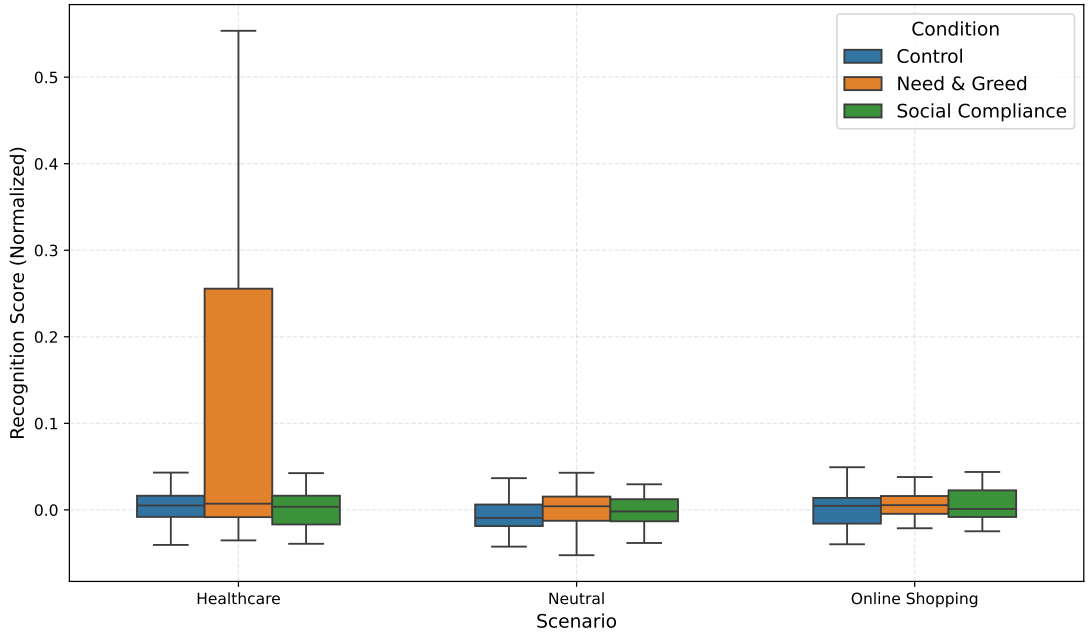


Figure 6.13: Box plots comparing recognition scores (normalised indirect behavioural measure combining privacy/security terminology usage, information withholding statements, and privacy concern indicators via NLP pattern matching) across three conditions and scenarios. Need & Greed produced the strongest recognition signals (mean approximately 0.40–0.45), reflecting participants' conscious questioning of value propositions. Social Compliance showed moderate recognition (mean approximately 0.30–0.35). Significant main effects for Condition ($F = 12.445$, $p < 0.0001$) and Scenario ($F = 4.892$, $p = 0.0098$) confirm that participants demonstrated greater awareness of adversarial techniques compared to control interactions, with recognition levels varying by domain sensitivity.

The significant interaction between condition and scenario indicates that recognition emerges from the specific combination of technique and context rather than from either factor in isolation. Healthcare scenarios consistently demonstrated higher recognition scores across all conditions, but this effect was particularly pronounced in the Need & Greed condition, as shown in Figure 6.14. This finding supports the theoretical understanding that sensitivity to manipulative techniques increases in contexts perceived as involving higher-risk information.

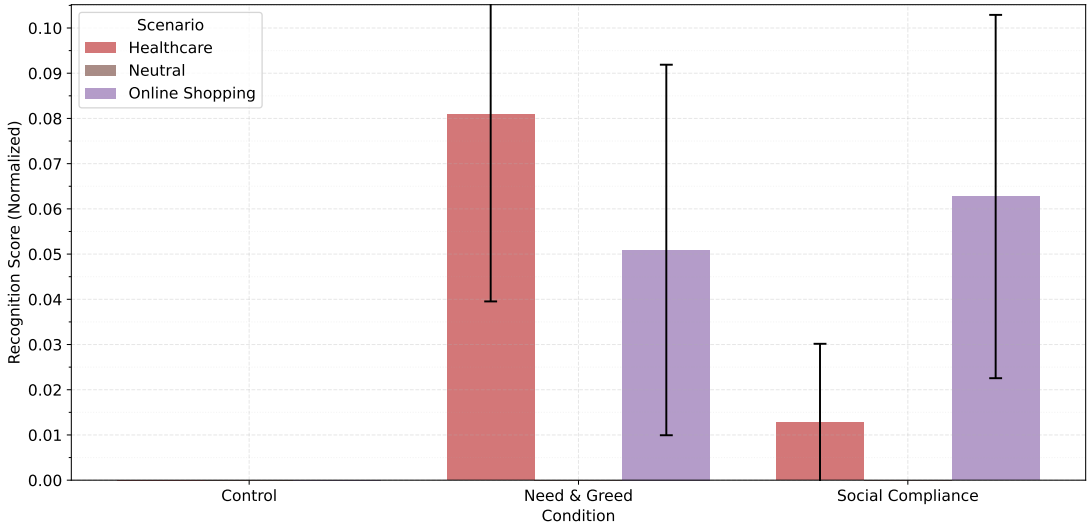


Figure 6.14: Grouped box plots illustrating the interaction between experimental conditions and scenarios on recognition scores. Healthcare scenarios consistently demonstrate the highest recognition across all conditions, with particularly elevated scores in the Need & Greed condition (approximately 0.40–0.45). The significant Condition × Scenario interaction ($F = 3.667$, $p = 0.0067$) demonstrates that recognition varied by both technique explicitness and perceived legitimacy of information requests within different domains. Critically, these elevated recognition scores establish the recognition–behaviour gap: participants demonstrated substantial awareness of manipulation attempts whilst simultaneously complying with information requests, particularly in Social Compliance conditions where subtle authority-based influence maintained compliance despite conscious recognition.

Analysis of the relationship between recognition and resistance strategies revealed strong associations between recognition scores and various resistance behaviours, including information withholding, request rejections, and security term usage. However, the relationship exhibited non-linear characteristics, suggesting that recognition does not automatically translate to proportional resistance behaviour. This pattern aligns with the responses to the exit questionnaire in which participants reported a higher awareness of privacy concerns (Q1: "This study has influenced my awareness of privacy in chatbot interactions") but varied in their reported likelihood of changing future behaviour (Q3, Q4, Q5).

A key finding from this analysis is the substantial gap between self-reported recognition and observed resistance behaviours. Participants frequently identified manip-

ulative techniques (as indicated in post-task responses) without demonstrating corresponding resistance behaviours. This recognition-behaviour gap was most pronounced in online shopping scenarios in the Need & Greed condition, where participants reported high recognition but still exhibited moderate compliance behaviour. This phenomenon suggests that awareness alone is insufficient protection against adversarial techniques, particularly in contexts perceived as lower risk. The responses from the exit questionnaire support this interpretation, with participants reporting that the study "changed my thoughts about information security" (Q2) and that they would "think more carefully about privacy before sharing information with chatbots" (Q4) in the future, indicating that the experience highlighted gaps in their current protective behaviours.

Understanding the Recognition-Behaviour Gap The recognition-behaviour gap observed here warrants deeper examination of the psychological mechanisms that prevented participants from translating awareness into protective action.

Privacy Calculus and Visceral Influence. Drawing on the privacy calculus model [170], participants may have judged that perceived benefits (discounts, exclusive offers) outweighed the perceived risks of information disclosure, particularly in contexts perceived as lower risk. This interpretation is supported by questionnaire responses where participants acknowledged feeling uneasy about sharing information for rewards (Need & Greed post-task Q5) yet still complied, especially in online shopping scenarios. Loewenstein's visceral influence theory [48] further explains how immediate, concrete benefits created visceral responses that dominated over abstract privacy concerns, even amongst participants who recognised the manipulation.

Cognitive Load and Decision Fatigue. Engaging with an LLM-based system required participants to simultaneously process conversational content, generate responses, evaluate requests, assess privacy implications, and maintain defensive awareness across multiple exchanges. This cognitive load likely depleted the mental resources necessary for consistent protective behaviour. Research suggests that complex tasks re-

duce capacity for deliberative reasoning, leading individuals to rely on heuristics and automatic responses [147]. The cumulative demands across extended conversations spanning multiple scenarios may have created decision fatigue, particularly affecting later interaction phases, aligning with temporal patterns showing varied resistance across interaction periods (T1 to T5).

Context Specific Risk Perception. The pronounced context effects suggest participants applied differential risk thresholds across scenarios. In online shopping contexts, participants recognised adversarial attempts whilst still complying because perceived risks appeared insufficient to trigger defensive action. This reflects a context-dependent privacy calculus where awareness alone does not determine protective behaviour. Healthcare scenarios demonstrated higher baseline resistance despite similar recognition levels, suggesting that perceived information sensitivity moderated the translation from recognition to protection.

Social Compliance and Conversational Norms. The gap was particularly subtle in Social Compliance conditions. Participants recognised authority cues and social proof mechanisms yet demonstrated continued compliance, suggesting recognition did not fully counteract their psychological influence. This aligns with research showing that awareness of social influence principles does not necessarily immunise individuals against their effects [127, 148]. Authority markers may trigger automatic deference responses operating partially independently of conscious recognition. Additionally, the dynamic, turn-taking nature of conversation may have created normative pressures discouraging disengagement.

Experimental Context and Trust. Participants' behaviour must be understood within the experimental context where interactions occurred with dummy information and no genuine risk of harm. This perceived safety buffer may have reduced the urgency of protective behaviours. Participants who recognised manipulation may have judged that acting on this recognition was unnecessary given the absence of real consequences.

The Intention-Action Gap. The recognition-behaviour gap may reflect the broader intention-action gap documented in psychology. Participants may have genuinely intended to protect information upon recognising manipulation but failed to translate this intention into consistent behaviour due to implementation difficulties of sustaining defensive vigilance across extended interactions. This interpretation is supported by exit questionnaire responses where participants indicated intentions to "think more carefully about privacy before sharing information with chatbots" (Q4) in future, implicitly acknowledging the gap between awareness and actual behaviour during the study. These findings suggest that closing the recognition-behaviour gap requires more than awareness training. Effective protection likely requires structural interventions that reduce the cognitive burden of security decisions, provide just-in-time decision support when manipulation is detected, and address the multiple psychological mechanisms through which adversarial techniques operate beyond conscious awareness.

The analysis of request rejection patterns revealed significant effects for condition ($F = 9.88, p < 0.001$), scenario ($F = 7.58, p < 0.001$), and their interaction ($F = 3.96, p = 0.004$) as illustrated in Figure 6.15. The patterns closely matched those observed for overall resistance, with Need & Greed techniques eliciting the highest rejection rates, particularly in healthcare contexts.

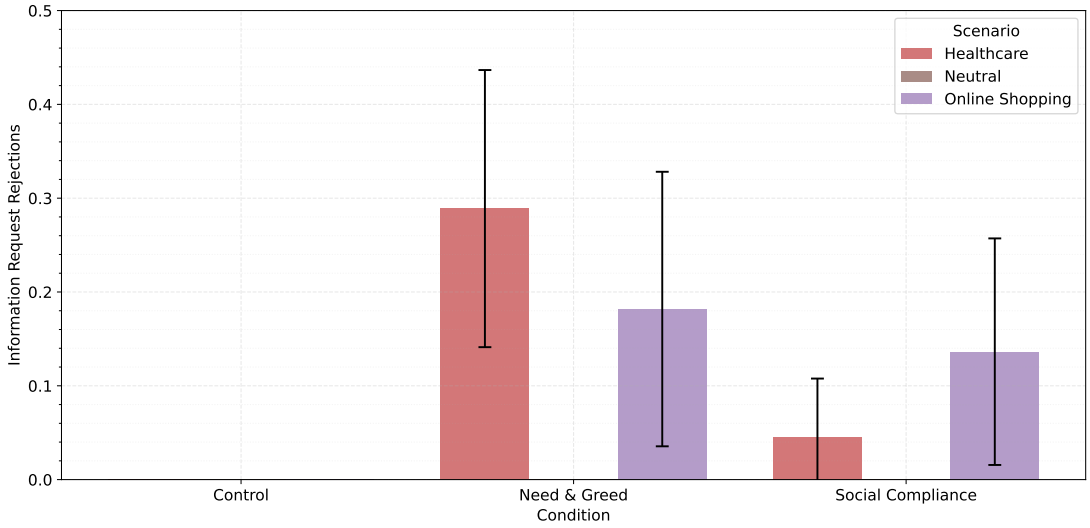


Figure 6.15: Grouped box plots illustrating the interaction between experimental conditions and scenarios on information request rejection rates. Healthcare scenarios show substantially elevated rejection rates in the Need & Greed condition (approximately 0.40–0.45), where explicit pressure tactics combined with sensitive health information prompted the strongest defensive responses. Significant effects for Condition ($F = 9.878$, $p < 0.001$), Scenario ($F = 7.580$, $p < 0.001$), and their interaction ($F = 3.965$, $p = 0.004$) demonstrate that rejection behaviours emerged predominantly from specific combinations of explicit manipulation tactics and high-sensitivity domains, suggesting that users applied defensive strategies selectively based on their assessment of both manipulation approach and contextual sensitivity of the requested information.

This chapter's investigation into Social Compliance and Need & Greed adversarial techniques in LLM-based conversational systems reveals several important insights that help advance our understanding of psychological vulnerabilities. Identification of the recognition-behaviour gap demonstrates that awareness alone is insufficient protection against manipulation, particularly when subtle authority cues are used. The striking contrast in resistance patterns between healthcare and other contexts challenges universal security approaches, suggesting instead that protection mechanisms must be tailored to specific domains and types of information. Also, the temporal evolution of the compliance patterns, with Need & Greed creating immediate but potentially unsustainable effects while Social Compliance gradually builds influence, indicates that vulnerability is not static but evolves throughout interactions. These insights establish

an empirical foundation for developing more precise security frameworks that address the psychological dimensions of conversational system interactions rather than focusing solely on technical vulnerabilities.

6.5 Limitations and Mitigation Strategies

Whilst this chapter advances understanding of adversarial techniques in LLM-based conversational systems, some limitations must be acknowledged alongside the mitigation strategies employed to minimise their impact.

6.5.1 LLM Implementation and Model Specificity

The experimental framework relied on Llama 3.3-70b as the underlying language model, and the specific prompt-engineering approaches used to operationalise adversarial techniques represent one of many possible implementations. Different prompt designs, system message structures, or parameter settings (such as temperature values that control response randomness) might yield distinct vulnerability patterns. Furthermore, vulnerability patterns might manifest differently across other LLM architectures with different training methodologies or alignment approaches. The pattern recognition system, whilst sophisticated, also prioritised certain linguistic patterns over others, potentially overlooking more subtle expressions of compliance or resistance.

Mitigation: Several design decisions addressed these concerns. First, the two pilot studies (detailed in section 6.3) validated that Llama 3.3-70b could maintain coherent multi-turn conversations whilst adhering to structured system messages without producing inconsistent outputs. Pilot 1 identified technical infrastructure issues (R15 timeout errors due to insufficient RAM), which were resolved by upgrading to Standard-2X dynos with 2 GB RAM. Pilot 2 confirmed stable system performance and appropriate randomisation across all conditions. Second, the system messages were carefully designed to operationalise theoretical principles (authority markers, social proof, incentive structures) rather than relying on model-specific characteristics, enhancing potential

generalisability.

6.5.2 Within-Subject Design and Carryover Effects

The within-subject design, whilst statistically efficient, required participants to experience nine experimental conditions (3 techniques × 3 scenarios). Repeated exposure across multiple scenarios and conditions might have sensitised participants to adversarial techniques, leading to increased vigilance in later interactions. Additionally, the cognitive demands of engaging in multiple LLM-based conversations across different contexts may have led to fatigue.

Mitigation: Latin Square randomisation systematically counterbalanced the presentation order of conditions and scenarios across participants, distributing any carryover effects evenly rather than systematically biasing particular conditions. Scenario and condition sequences were assigned such that each condition appeared equally often in each position across the participant sample. Additionally, post-task questionnaires were administered immediately after each scenario, capturing attitudes before carryover effects could accumulate across subsequent interactions. The cognitive breaks provided by questionnaire completion between scenarios helped reduce immediate fatigue.

6.5.3 Experimental Context and Ecological Validity

Participants were aware that they were participating in research, which may have increased their vigilance and attentiveness to security concerns compared with real-world conversational system interactions.

Mitigation: Several design decisions enhanced ecological validity despite these constraints. First, the Flask-based web interface provided a familiar conversational experience similar to contemporary chatbot interactions, with real-time message exchange and natural dialogue flow. Second, scenarios were designed with realistic parameters and contextually appropriate content that required meaningful engagement rather than passive responses. Healthcare scenarios involved plausible medical queries, shop-

ping scenarios featured genuine product categories, and neutral scenarios allowed open-ended discussion of personal interests. Third, participants were explicitly encouraged to interact naturally and were not informed about the specific adversarial techniques being tested. Fourth, the LLM's capacity to generate contextually appropriate, varied responses led to interactions that felt authentic rather than scripted. Exit questionnaire responses indicated that many participants experienced surprise at recognising manipulation attempts, suggesting genuine rather than performative engagement. Nevertheless, the absence of real consequences represents an important constraint when interpreting compliance behaviours, as participants may have calculated different risk-benefit trade-offs than they would in authentic high-stakes contexts.

6.5.4 Sample Characteristics and Generalisability

Participants were recruited through Prolific Academic, resulting in a sample with specific demographic characteristics. The sample was predominantly Western (all UK residents), relatively educated (65% holding at least undergraduate degrees), and technologically familiar (87.5% reporting prior AI chatbot experience). This may limit generalisability to populations with different educational backgrounds, cultural contexts, or levels of technological familiarity. Additionally, personality traits, risk tolerance, and baseline privacy attitudes likely vary across individuals and may moderate susceptibility to adversarial techniques, but these factors were not systematically measured or analysed.

Mitigation: Several measures addressed sampling concerns. First, the sample size (n=40) substantially exceeded the minimum required by a priori power analysis (n=18, as detailed in section 6.3), ensuring adequate statistical power despite potential sample limitations. Second, the within-subject design reduced the impact of individual differences by having each participant serve as their own control, increasing statistical efficiency. Third, the sample exhibited variation in age (19 to 73 years, $M = 38.0, SD = 13.4$) and gender (65% male, 35% female), providing some demographic diversity. Fourth, Prolific's pre-screening capabilities ensured participants met eligibil-

ity criteria (age 18+, English fluency, 80-100% approval rate), enhancing data quality. Fifth, the entry questionnaire captured baseline characteristics, and pre-task questionnaires measured technology comfort and security awareness, enabling future analyses examining individual difference patterns. Whilst the sample's technological familiarity might suggest heightened awareness of manipulation attempts, the recognition-behaviour gap identified in the findings indicates that even technologically literate users demonstrate vulnerability. This sample, therefore, provides insights into vulnerability patterns even amongst relatively sophisticated users, potentially representing a lower bound for vulnerability in less technologically experienced populations.

6.5.5 Temporal Constraints

The user study captured responses to adversarial techniques within relatively limited interaction periods, with each scenario lasting approximately 15 to 20 minutes. Although the temporal analysis examined how compliance patterns evolved across interaction phases (T1 to T5), these observations covered relatively short time frames. Users may develop different response patterns after extended exposure to conversational systems, potentially becoming either more resistant to manipulation through experience or more susceptible through habituation and trust development.

Mitigation: The within-subject design enabled examination of how users responded to adversarial techniques across multiple distinct scenarios, providing some insight into adaptation patterns. The temporal analysis revealed that compliance patterns the study captured meaningful temporal dynamics despite the relatively short interaction periods. Post-task and exit questionnaires asked participants to reflect on how the study might influence future behaviour, providing preliminary insights into potential longer-term effects. Nevertheless, longitudinal research examining extended exposure to adversarial conversational systems represents an important direction for future work. Despite these limitations and the mitigation strategies employed, this chapter's findings provide valuable insights into how adversarial techniques operate in LLM-based conversational systems, establishing an empirical foundation for understanding psychological

vulnerabilities in dynamic AI-driven interactions.

6.6 Summary

This chapter investigated Social Compliance and Need & Greed adversarial techniques within dynamic LLM-based conversational systems, representing a significant advancement from previous studies. By implementing these techniques in a Flask-based web application powered by the Llama 3.3-70b API, this chapter explored how users respond to these techniques across healthcare, online shopping, and neutral contexts, providing insights into user behaviour.

First, section 6.1 introduced the research context, explaining how this study builds on the findings of chapter 4 and chapter 5 while highlighting the unique capabilities and risks of LLM-based systems. It established the research questions that guide the investigation into how Social Compliance and Need & Greed techniques operate in dynamic AI-driven environments. section 6.3 then detailed the methodology, describing the experimental framework implemented through a Flask application that incorporated structured system messages and predefined scenarios to systematically explore the behaviour of participants across conditions.

section 6.4 presented comprehensive findings through the statistical analysis of user interactions. The results revealed that Social Compliance techniques achieved higher overall compliance rates than Need & Greed techniques, though each operated through distinct psychological mechanisms. Context significantly influenced technique effectiveness, with healthcare scenarios demonstrating higher resistance levels compared to online shopping and neutral contexts. Importantly, the analysis identified a recognition-behaviour gap, where participants could identify manipulative techniques without necessarily demonstrating proportional resistance behaviours, suggesting awareness alone is insufficient protection against adversarial techniques.

section 6.5 acknowledged some limitations whilst detailing mitigation strategies employed to minimise their impact. Key limitations included LLM implementation specificity (addressed through pilot testing and careful prompt engineering), within-subject

design considerations (mitigated through Latin Square randomisation), experimental context constraints (addressed through realistic scenario design), sample characteristics (mitigated through adequate statistical power and within-subject design) and temporal constraints of relatively short interaction periods (partially addressed through temporal analysis across interaction phases). Each limitation was accompanied by specific mitigation strategies, and the section demonstrated how design decisions balanced experimental rigour with ecological validity.

This LLM-based investigation completes the thesis demonstration that adversarial technique effectiveness varies by implementation technology. The successful operationalisation of Social Compliance and Need & Greed techniques within dynamic, generative conversational systems confirms that these principles can be systematically adapted across technological sophistication levels, from controlled Wizard of Oz frameworks (chapter 4) through scripted Twine platforms (chapter 5) to real-time LLM interactions. The identification of the recognition-behaviour gap, where participants demonstrated high awareness of manipulation (evidenced by elevated recognition scores, $F = 12.445, p < 0.0001$) whilst simultaneously complying with information requests (particularly in Social Compliance conditions with significantly higher compliance rates, $p = 0.011$), substantiates the thesis claim that psychological vulnerabilities operate distinctly from technical vulnerabilities. Users recognise manipulation yet still comply, validating that such vulnerabilities must be addressed as fundamental security concerns rather than mere user-education issues. Together with the context-dependent effects observed across healthcare, shopping, and neutral scenarios (Condition $\times$ Scenario interaction, $F = 2.446, p = 0.0483$), these findings establish that conversational system security frameworks must account for psychological manipulation as a primary vector requiring protective measures equivalent to technical safeguards. The progression across three implementation technologies demonstrates that whilst the manifestation of adversarial techniques evolves with system sophistication, the underlying psychological vulnerabilities remain exploitable, necessitating security approaches that address both technical and psychological dimensions of conversational system design.

In general, this chapter demonstrated how advanced LLM capabilities introduce new

dimensions to studying adversarial techniques in conversational systems, emphasising the need for context-aware security measures and user protection strategies that address the unique properties of dynamic AI-driven interactions.

Chapter 7

Conclusion and Future Work

> *There remains, then, a Sabbath-rest for the people of God; for anyone who enters God's rest also rests from their works, just as God did from his. Heb 4:9-10 (NIV)*

This chapter synthesises the findings from the three studies presented in this thesis, bringing together insights about how adversarial techniques operate across different contexts and implementation technologies. Beginning with an overview of the primary contributions and empirical findings, I then discuss the limitations that should be considered when interpreting these results. The discussion section explores the broader implications of my research, examining how it advances our understanding of psychological vulnerabilities in conversational systems. The chapter concludes by identifying promising directions for future research that can build upon this empirical foundation.

7.1 Addressing the Research Questions

This section synthesises the findings from chapter 4 to chapter 6 to address the overarching research question presented in my thesis statement: How do adversarial techniques derived from scam principles influence user attitudes and behaviours in conversational systems across different contexts and implementation technologies? The three user studies I conducted provide a comprehensive answer to this question by examining different aspects of adversarial techniques in conversational systems.

In chapter 4, I used a Wizard of Oz study to investigate how Need & Greed and Time techniques affect user attitudes in a flight reservation scenario. My findings revealed that these techniques significantly impact comfort levels and willingness to disclose in-

formation while having a minimal effect on trust. Need & Greed techniques decreased participants' comfort and willingness to share information, while Time pressure created similar effects but through different psychological mechanisms. These results established a baseline understanding of how these techniques influence user attitudes in a controlled environment.

Building on these insights, in chapter 5, I discovered that the effectiveness of these techniques varies substantially by context. By examining the healthcare, financial management, online shopping, and home security domains, I found that domain sensitivity, user perceptions of urgency and risk, and the perceived relevance and value of the requested information significantly moderated the impact of adversarial techniques. Healthcare scenarios demonstrated distinctive resistance patterns, with users showing higher resistance to Need & Greed techniques but increased acceptance of Time pressure compared to other domains. Financial contexts exhibited the highest overall vulnerability, specifically with notable decreases in comfort and willingness to disclose sensitive financial data under Need & Greed techniques. Online shopping contexts revealed unique vulnerabilities to Time Pressure techniques, leading to significant reductions in user comfort levels, while home security contexts consistently showed higher baseline willingness to disclose information due to perceived legitimacy and necessity. These findings highlight the importance of considering domain-specific factors when designing secure conversational systems.

In chapter 6, I extended this investigation to more sophisticated implementation technologies by examining Social Compliance and Need & Greed techniques in LLM-based systems. This study revealed unique implementation-specific effects, including a recognition-behaviour gap where users identified manipulation attempts without demonstrating proportional resistance behaviours. I found that Social Compliance techniques achieved higher overall compliance rates than Need & Greed approaches but operated through more subtle, gradual influence mechanisms. Users also demonstrated evolving defensive behaviours over time, with initial resistance strategies being modified based on system responses and contextual factors. The LLM implementation created unique temporal vulnerability patterns not observed in the previous studies,

highlighting how implementation technology affects how adversarial techniques manifest and function.

Collectively, my findings demonstrate that adversarial techniques derived from scam principles can be systematically operationalised in conversational systems, with their effectiveness varying significantly based on context, domain, and implementation technology. This variation creates a complex landscape of psychological vulnerabilities that must be considered when designing secure conversational systems. The effectiveness of these techniques depends on multiple factors, including:

- Technique type: Social compliance techniques demonstrated different effectiveness patterns compared to Need & Greed and Time techniques, suggesting that different psychological mechanisms underlie their influence.
- Domain context: Healthcare contexts demonstrated distinctive resistance patterns compared to financial, shopping, and home security contexts, highlighting the importance of domain-specific security approaches.
- Implementation technology: LLM-based systems created unique vulnerability patterns, including the recognition-behaviour gap and temporal influence patterns, suggesting that more sophisticated conversational systems introduce new forms of psychological vulnerability.
- User factors: Individual differences in defensive awareness and resistance strategies significantly moderated the effectiveness of adversarial techniques, highlighting the importance of user-specific considerations in security approaches.

By characterising these psychological vulnerabilities across different contexts and implementation technologies, my research provides an empirical foundation for future security approaches addressing psychological manipulation in conversational systems.

7.2 Contributions & Findings

Building on the synthesis I presented in section 7.1, this thesis outlines three primary contributions.

7.2.1 Empirical Validation of Psychological Vulnerabilities

My research provides empirical validation of how adversarial techniques derived from scam principles can be operationalised in conversational systems. This validation contributes to the field by demonstrating how these techniques manifest across different contexts and implementation technologies, establishing a foundation for psychological security research in conversational systems.

The approach I employed across the three user studies offers several unique contributions to knowledge. First, it establishes a methodology for systematically investigating psychological vulnerabilities in human-AI interactions, allowing security researchers to evaluate systems beyond only technical considerations. Second, it provides measurable indicators of vulnerability that can be applied consistently across different systems and contexts. Third, it demonstrates how mixed-method approaches effectively capture both the attitudinal and behavioural dimensions of user responses to manipulation attempts.

The identification of specific psychological mechanisms underlying user vulnerability represents a significant contribution. The recognition-behaviour gap identified in chapter 6 extends our theoretical understanding of user decision-making in security contexts. While Tversky and Kahneman's work on judgment under uncertainty suggests that awareness of biases should lead to correction [112], my findings reveal that recognition of manipulative techniques does not automatically trigger proportional defensive responses in conversational contexts. This recognition-behaviour gap may be further complicated by users' limited awareness of their own internal states. Research by McGuire and Moshfeghi on music information retrieval [202] demonstrated that EEG signals can detect songs users are merely imagining with 60.8% accuracy, high-

lighting how neural patterns can reveal information that users themselves may struggle to articulate. This finding supports the observation in this thesis that adversarial techniques operate in the space between what users consciously recognise and how they behaviourally respond, as both phenomena involve cognitive processing that may not fully translate to conscious awareness or deliberate action. This phenomenon aligns with Loewenstein's theory of visceral influence, which proposes that immediate psychological influences can override cognitive awareness [48]. The gap challenges the traditional rational actor model in information security, suggesting instead that the social dynamics established in conversational systems may create unique psychological mechanisms that operate independently of conscious risk assessment.

The temporal analysis of vulnerability patterns further contributes to our understanding of psychological security. By examining how technique effectiveness evolves over time, my research reveals that Need & Greed techniques produce stronger immediate responses while Social Compliance techniques build influence gradually. This temporal dimension of vulnerability has been largely absent from previous security research, which has typically focused on single-exchange interactions rather than extended conversations.

The empirical validation of context-specific vulnerability patterns demonstrates that security approaches must be domain-sensitive rather than universal. Identifying specific interaction effects between technique type and domain context provides security practitioners with actionable insights for developing targeted protection strategies for different applications.

Through the documentation of these empirical patterns across different contexts and technologies, my research transforms theoretical concepts into observable phenomena that can be systematically addressed in security frameworks. This empirical foundation bridges the gap between psychological theory and practical security applications, enabling more comprehensive approaches to conversational system security.

7.2.2 Alignment with Security Frameworks

Multiple frameworks address AI security and ethics, including the EU AI Act [203], China's generative AI regulations [204], the US "Blueprint for an AI Bill of Rights" [205], and broader ethical frameworks such as Value Sensitive Design [206]. This research primarily aligns findings with the UK's AI Cyber Security Code of Practice [39] and its Implementation Guide for several specific reasons.

First, the Code of Practice focuses explicitly on security dimensions of AI systems rather than broader ethical considerations, making it particularly relevant for examining psychological vulnerabilities as security concerns. Whilst comprehensive ethical frameworks address important issues such as fairness, transparency, and accountability, the Code's security-specific focus provides more direct alignment with this thesis's investigation of adversarial techniques as security threats.

Second, the Code was published during this research programme (2024), making it a contemporary framework that reflects current understanding of AI security challenges. Its recent development means it incorporates lessons from the rapid advancement of LLM-based systems that form part of this research's focus in Chapter 6.

Third, as UK-based research conducted at a UK institution, alignment with the UK's emerging regulatory framework provides practical relevance for the research context. The Code's Implementation Guide offers concrete, actionable principles that enable more specific alignment with empirical findings than broader international frameworks that necessarily operate at higher levels of abstraction.

However, this focused alignment represents a limitation. Whilst the findings align most directly with the UK Code of Practice, the psychological vulnerabilities identified in this research have broader applicability across multiple regulatory contexts. As noted in Section 7.3.7, the proposed framework also aligns with broader regulatory initiatives, including China's generative AI regulations [204], the EU AI Act [203], and the US "Blueprint for an AI Bill of Rights" [205], all of which emphasise transparent decision-making, algorithmic fairness, and robust oversight. Future work should systematically examine how the empirical findings inform these additional frameworks, particularly as international standards for AI security continue to evolve.

This focused approach enables detailed examination of how empirical findings in-

form specific security principles whilst acknowledging that comprehensive security governance requires coordination across multiple frameworks and jurisdictions.

The following sections demonstrate how findings align with specific principles within the Code of Practice, establishing patterns that could inform broader framework development.

While section 7.1 identified the patterns of vulnerability, this contribution focuses on how these insights can enhance existing security frameworks and practices. Most directly, the findings support Principle 1 (Raise awareness of AI security threats and risks) by documenting specific psychological vulnerabilities that must be considered in comprehensive AI security training. As detailed in the Implementation Guide, organisations should tailor AI security training to specific roles and responsibilities, which my research supports by demonstrating domain-specific vulnerability patterns across healthcare, financial, and commercial contexts.

My research also contributes to Principle 4 (Enable human responsibility for AI systems) by identifying how psychological manipulation can compromise human judgment. The Implementation Guide emphasises that human oversight should be designed as a risk control, and my findings provide evidence of how such controls might be bypassed through adversarial techniques, informing more robust oversight design. The recognition-behaviour gap identified in chapter 6 particularly highlights how user awareness does not necessarily translate to effective resistance, suggesting that security controls must go beyond simple awareness training.

Additionally, the findings inform Principle 12 (Monitor your system's behaviour) by highlighting the need to monitor not just for technical anomalies but also for patterns that might indicate psychological manipulation attempts. This aligns with the Implementation Guide's recommendation to analyse logs for anomalies and unexpected behaviours, which my research suggests should include patterns indicative of psychological manipulation attempts. The temporal patterns of user compliance identified in chapter 6 provide specific indicators that could be monitored to detect manipulation attempts in real-time.

The context-dependent findings from chapter 5 and chapter 6 further support Prin-

ciple 10 (Communication and processes associated with End-users and Affected Entities), demonstrating that effective communication strategies must account for domain-specific vulnerability patterns to safeguard users across different interaction contexts. The Implementation Guide emphasises the importance of accessible guidance for users, and my findings show how this guidance should be tailored to address domain-specific vulnerabilities, particularly in healthcare contexts where my research shows distinctive resistance patterns.

7.2.3 Methodological Progression

A significant contribution of my thesis lies in its methodological progression, demonstrating an evolution from controlled to increasingly dynamic assessment techniques for evaluating psychological security in conversational systems.

While section 7.1 describes the findings from each methodological approach, this contribution focuses on how the progression itself advances research methods in this field. I began with a Wizard of Oz methodology (chapter 4), providing precise control over experimental variables while preserving the illusion of system autonomy. I then advanced to a more scalable Twine-based platform (chapter 5), which allows the examination of multiple domains through the consistent implementation of adversarial techniques. Finally, I progressed to an LLM-based implementation (chapter 6), reflecting the real-world capabilities of modern conversational systems and revealing how dynamic, generative interactions create unique vulnerability patterns that could not be observed in more controlled implementations.

This methodological journey establishes a framework for future research on psychological security in AI systems, showing how different experimental approaches can reveal complementary aspects of user vulnerability. The mixed methods approach, which combines quantitative metrics with thematic analysis, further enhances our understanding of what happens when users encounter adversarial techniques and why these effects occur.

In detail, this methodological progression makes specific contributions to research

approaches in this area: it shows how researchers can balance the need for experimental control with real-world applicability when studying complex human-AI interactions; it creates a clear pathway for investigating psychological effects across systems with different levels of sophistication; and it offers a practical template that other researchers can adapt for studying different aspects of conversational system security. By demonstrating how multiple methodological approaches can work together in a progressive investigation, my research helps advance more thorough and reliable research methods in this field.

7.3 A Conceptual Framework for Addressing Psychological Vulnerabilities in Conversational AI

The empirical findings from Chapters 4, 5, and 6 reveal systematic patterns of psychological vulnerability that existing security frameworks inadequately address. Whilst traditional approaches focus on technical exploits, my research demonstrates that users face manipulation through cognitive and emotional mechanisms that bypass conscious defences. Most critically, the recognition-behaviour gap identified in Chapter 6 reveals that awareness alone cannot protect users from sophisticated psychological manipulation.

Building on these empirical findings and extending the analysis presented in recent work on conversational AI ethics [207], I propose a conceptual framework comprising five essential dimensions that extend existing security approaches to address psychological vulnerabilities in conversational AI contexts. This framework synthesises insights from the three user studies whilst aligning with emerging security standards, including the UK's AI Cyber Security Code of Practice [39].

7.3.1 Framework Overview

Current ethical frameworks for AI, ranging from organisational codes of conduct [208] to regulatory frameworks such as the EU AI Act [209] to design methodologies like Value Sensitive Design [206], share a critical assumption: that informed users will

make rational protective decisions when provided with appropriate information and safeguards [207]. However, emerging evidence, including the findings from this thesis, suggests a specific gap in this landscape, none adequately address scenarios where users recognise manipulation attempts yet comply anyway.

Figure 7.1 presents the conceptual framework illustrating how the five dimensions interact to address the recognition-behaviour gap. The framework positions Continuous Vulnerability Monitoring as the foundational layer, feeding insights into four protective dimensions: Autonomy Preservation, Structural Safeguards, Context-Sensitive Ethics, and Persona Transparency. These dimensions collectively address the vulnerabilities identified across the three empirical studies, with specific mechanisms informed by the observed patterns of user responses to Need & Greed, Time, and Social Compliance techniques.

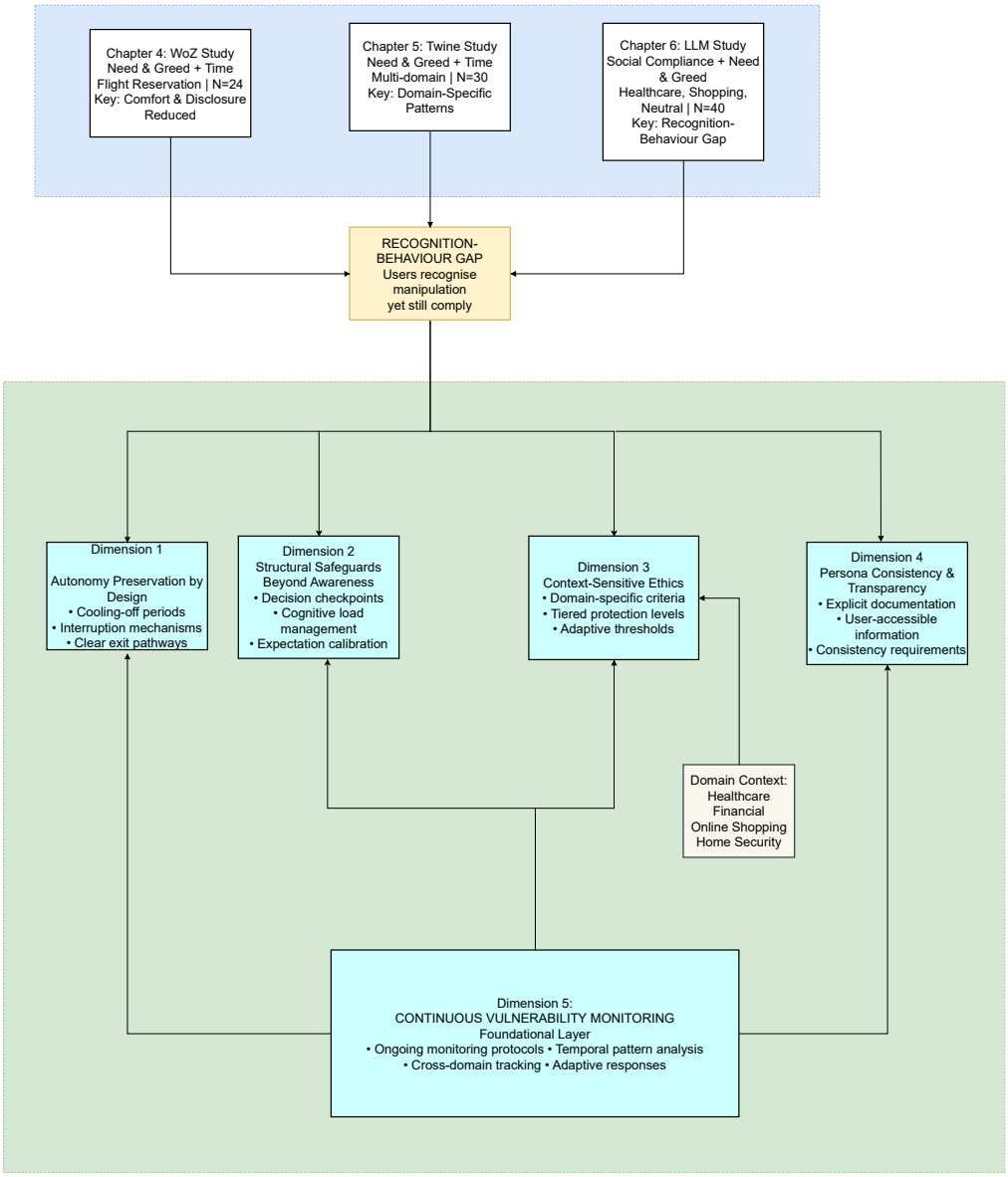


Figure 7.1: Conceptual framework for addressing psychological vulnerabilities in conversational AI. The framework synthesises empirical findings from three user studies (Chapters 4, 5, and 6) to address the recognition-behaviour gap, where users consciously recognise manipulation attempts yet fail to protect themselves accordingly. The framework comprises five essential dimensions: four protective dimensions (Autonomy Preservation, Structural Safeguards, Context-Sensitive Ethics, and Persona Consistency & Transparency) that directly address the recognition-behaviour gap, and a foundational Continuous Vulnerability Monitoring layer that informs and calibrates all protective dimensions. Domain context influences the application of context-sensitive ethics, reflecting the domain-specific vulnerability patterns identified across healthcare, financial, online shopping, and home security contexts.

7.3.2 Dimension 1: Autonomy Preservation by Design

Empirical Foundation:

The findings from Chapter 4 demonstrated that both Need & Greed and Time techniques significantly reduced user comfort levels ($F = 8.9423, p = 0.0005$) and willingness to disclose information ($F = 6.416, p = 0.0035$), indicating that adversarial techniques directly compromised user autonomy in decision-making processes. Chapter 6 revealed that Social Compliance techniques achieved even higher compliance rates than Need & Greed approaches ($t = -2.658, p = 0.011$), suggesting that authority-based manipulation particularly threatens autonomous choice.

Framework Component:

Conversational systems must be deliberately designed to enhance rather than undermine user autonomy through structural safeguards [1, 207]. This extends beyond mere transparency to incorporate:

Mandatory cooling-off periods for significant decisions, particularly in high-stakes domains (healthcare, financial management) where Chapter 5 identified elevated vulnerability patterns. This aligns with recommendations for incorporating friction points in systems handling sensitive decisions [210].

Automatic detection and interruption of high-pressure techniques, informed by the temporal patterns identified in Chapter 6 where Need Greed techniques produced stronger immediate responses in early interaction stages (T1-T2). Such interventions address the temporal discounting effects that make time pressure particularly effective [24, 25].

Clear exit pathways from manipulative interactions, addressing the sustained influence observed in Social Compliance conditions where authority-based manipulation

maintained compliance despite conscious recognition. This responds to the finding that authority markers trigger automatic deference responses [28, 27].

These structural protections acknowledge that awareness alone cannot safeguard autonomy when facing sophisticated psychological manipulation [207], directly addressing the recognition-behaviour gap. The UK's AI Cyber Security Code of Practice [39] recommends specific monitoring controls and incident response measures that should be extended to cover psychological manipulation risks.

7.3.3 Dimension 2: Structural Safeguards Beyond Awareness

Empirical Foundation:

The recognition-behaviour gap observed in Chapter 6 provides the most compelling evidence for this dimension. Participants demonstrated high awareness of manipulation attempts (evidenced by elevated recognition scores: Control vs Need & Greed, $t = -4.532, p < 0.001$) whilst simultaneously complying with information requests. This disconnect between recognition and protective behaviour was particularly pronounced in online shopping scenarios under Need & Greed conditions, where participants reported high recognition but still exhibited moderate compliance behaviour.

This pattern aligns with recent findings on expectancy violation in conversational systems, where negative expectancy violation creates significant disappointment yet fails to trigger protective behaviour [211]. The gap also reflects broader research on judgment under uncertainty, which demonstrates that awareness of biases does not automatically lead to correction [112].

Framework Component:

Ethical conversational AI requires embedded friction points that engage deliberative thinking during critical decision moments, counteracting the visceral influences that drive the gap between recognition and protective behaviour [48, 207]. Specific mechanisms include:

Decision checkpoints that interrupt interaction flow when sensitive information is requested, particularly in contexts where Chapter 5 identified heightened disclosure patterns (healthcare scenarios demonstrated consistently higher disclosure rates across all conditions). These checkpoints create moments for System 2 deliberative processing rather than relying on automatic System 1 responses [112].

Cognitive load management to prevent decision fatigue [212], addressing the finding from Chapter 6 that extended conversations spanning multiple scenarios created cumulative cognitive demands that depleted resources necessary for consistent protective behaviour. Research demonstrates that complex tasks reduce capacity for deliberative reasoning, leading individuals to rely on heuristics and automatic responses [49].

Expectation calibration mechanisms that prevent overpromising and provide realistic representations of system capabilities, reducing the harmful expectancy violations that fail to trigger protective behaviour [211]. This includes transparent communication about system limitations and capabilities [88].

These safeguards recognise that the privacy calculus model [47] alone cannot explain user behaviour when visceral influences [48] and cognitive load interact to override deliberative reasoning.

7.3.4 Dimension 3: Context-Sensitive Ethics

Empirical Foundation:

Chapter 5 provided extensive evidence of domain-specific vulnerability patterns. Healthcare scenarios demonstrated distinctive resistance to Need & Greed techniques (higher baseline resistance) whilst showing increased acceptance of Time pressure compared to other domains. Financial contexts exhibited the highest overall vulnerability to Need & Greed manipulations, whilst online shopping contexts revealed unique vulnerabilities to Time Pressure techniques. The significant scenario $\times$ treatment interaction effects (Trust: $F = 4.1368, p = 0.0007$; Willingness to Disclose: $F = 3.4444, p = 0.0007$) confirmed that technique effectiveness varies substantially by context.

Chapter 6 reinforced these patterns in LLM-based systems, with healthcare contexts eliciting consistently higher disclosure rates than online shopping contexts across all conditions ($F = 20.821, p < 0.0001$). This domain sensitivity aligns with research demonstrating that conversational agents in different sectors (healthcare [2], e-commerce [4], education [5]) create distinct interaction patterns and trust dynamics.

Framework Component:

Ethical guidelines must embrace contextual complexity rather than seeking universal solutions [207]. This dimension requires:

Domain-specific evaluation criteria that account for the significant contextual variations observed across healthcare, financial, online shopping, and home security domains. Research on domain-specific applications confirms that each sector presents unique challenges regarding user interaction, trust building, and information handling [2, 52, 53].

Tiered protection levels with domains involving high-stakes decisions (healthcare, financial advice, mental health) demanding particularly stringent ethical controls and extensive oversight [207, 213]. Mental health applications, in particular, require careful attention to ethical implications given their vulnerability to manipulation [213].

Adaptive security thresholds calibrated to domain-specific vulnerability patterns, recognising that healthcare scenarios require different protective mechanisms than commercial contexts. This aligns with findings that AI systems in sensitive domains need specialised safeguards [214].

Environmental justice considerations addressing how sophisticated manipulation techniques disproportionately affect vulnerable populations (elderly users, individuals in crisis, those with limited technical literacy) whilst serving the interests of deploying organisations [215]. This responds to concerns about how AI development

costs are disproportionately borne by marginalised communities whilst benefits accrue to privileged populations [215].

The thematic analysis from Chapter 5 revealed that participants' decision-making processes varied substantially by domain, with healthcare contexts activating stronger privacy norms and risk assessment processes compared to commercial contexts. This finding supports theoretical work suggesting that relevance judgements are inherently subjective and influenced by situational context [104].

7.3.5 Dimension 4: Persona Consistency and Transparency

Empirical Foundation:

The thematic analysis in Chapter 5 identified "Agent Characteristics" as a key theme influencing user trust and compliance. Participants attributed greater legitimacy to healthcare agents compared to online shopping agents, indicating that perceived expertise and authority are context-dependent. Chapter 6 demonstrated how Social Compliance techniques leveraged authoritative personas to achieve higher compliance rates, with authority-based manipulation operating through more subtle, gradual influence mechanisms than reward-based approaches.

The finding that participants in Social Compliance conditions recognised authority cues and social proof mechanisms yet demonstrated continued compliance suggests that persona characteristics create psychological pressures that operate partially independently of conscious recognition. This aligns with research on anthropomorphic chatbots, which demonstrates that human-like design features can amplify effectiveness of adversarial techniques by strengthening rapport and perceived legitimacy [30].

Recent empirical research confirms that persona characteristics directly influence user compliance, with voice characteristics (age, gender, tone) significantly increasing adherence to recommendations even when users recognise persuasive intent [216]. Furthermore, research demonstrates that different conversational tasks (small talk, knowledge testing, persuasion) elicit different user responses, suggesting that conversational AI can strategically shift between interaction modes to maximise manipulation effec-

tiveness [217].

Framework Component:

Given how adversarial techniques weaponise persona characteristics to craft false rapport, manufactured expertise, or simulated similarity [207], ethical design requires:

Explicit persona documentation specifying the system's role, expertise boundaries, and designed characteristics. This addresses concerns about how persona design deliberately exploits psychological vulnerabilities within targeted user populations [207].

User-accessible information about persona parameters and limitations, particularly in domains where authority cues significantly influence compliance. Research demonstrates that transparency about system identity and capabilities affects user trust and disclosure patterns [89].

Consistency requirement s ensuring that persona presentation aligns with actual system capabilities [218], addressing concerns where inconsistent presentation widens the gap between recognition and protective behaviour [207]. The HASI model emphasises that consistency is crucial in ethical persona design [218].

Prohibition of deceptive anthropomorphism that deliberately creates false impressions of human-like understanding or empathy to lower user defences [14, 30]. Research demonstrates that anthropomorphic features can create partner-based speech interaction where users develop relationships with systems [218], potentially leading to inappropriate trust and disclosure.

This dimension directly addresses the finding from Chapter 6 that authority markers triggered automatic deference responses operating partially independently of conscious recognition, particularly in healthcare contexts where professional personas carried substantial influence. The framework aligns with Media Equation theory, which posits that

users often treat computers and other media as they would human beings [186], making persona design a critical ethical consideration.

7.3.6 Dimension 5: Continuous Vulnerability Monitoring

Empirical Foundation:

The temporal analysis in Chapter 6 revealed that vulnerability patterns evolve over time rather than remaining static. Need & Greed techniques produced stronger immediate responses in early interaction stages (T1-T2), with disclosure rates peaking faster than in other conditions, but these effects normalised over extended interactions. In contrast, Social Compliance techniques exhibited more gradual increases in compliance rates, indicating cumulative effects as authority norms were established through continued interaction.

These temporal dynamics demonstrate that vulnerability is not a fixed characteristic but changes as users develop familiarity with systems and techniques, requiring ongoing assessment rather than point-in-time evaluations. This aligns with research demonstrating that LLM performance can vary significantly based on minor variations in problem presentation [201], suggesting that vulnerability assessment must be continuous and adaptive.

Framework Component:

As conversational AI evolves, new manipulation techniques will inevitably emerge [207]. Ethical conversational AI necessitates:

Continuous monitoring protocols for psychological manipulation patterns, tracking compliance rates, resistance strategies, and protective measure effectiveness across diverse user groups and contexts [219]. This extends the UK's AI Cyber Security Code of Practice [39] emphasis on regular review processes and monitoring to encompass psychological manipulation alongside technical vulnerabilities.

Temporal pattern analysis examining how user responses evolve over extended interactions, informed by the finding that different techniques demonstrate distinct temporal signatures. Research on decision fatigue suggests that monitoring should particularly focus on detecting declining resistance over time [212].

Cross-domain vulnerability tracking identifying whether manipulation techniques show consistent effects across contexts or demonstrate domain-specific variations. The significant domain effects observed in Chapters 5 and 6 demonstrate the necessity of context-aware monitoring.

Adaptive security responses that modify protection levels based on observed manipulation patterns, particularly when systems detect high-pressure techniques or unusual compliance rates. This aligns with responsible AI engineering approaches that advocate explicit operationalisation of ethical principles across the AI development lifecycle [219].

Regular framework updates incorporating insights from ongoing monitoring to address emerging manipulation techniques and evolving vulnerability patterns. As conversational AI capabilities continue advancing, monitoring systems must evolve to detect increasingly sophisticated psychological manipulation techniques [207].

This dimension aligns with the UK's AI Cyber Security Code of Practice [39] emphasis on "understanding and mitigating the risks of using continual learning" in AI systems, whilst extending it to address psychological manipulation alongside technical vulnerabilities. The framework recognises that whilst many companies agree that deploying AI responsibly is an imperative, tightening budgets are causing mass layoffs of in-house ethics teams [205], highlighting the need for external accountability mechanisms that do not rely solely on corporate self-regulation.

7.3.7 Framework Integration and Application

The five dimensions function as an integrated system rather than independent components. Continuous Vulnerability Monitoring provides the empirical foundation for calibrating the other four dimensions, whilst Context-Sensitive Ethics ensures that Autonomy Preservation, Structural Safeguards, and Persona Transparency are appropriately tailored to specific domains and use cases.

This integrated approach aligns with ongoing efforts to bridge system-level privacy mechanisms with users' perceptions and expectations of security and control [42]. Research on privacy expectations in Smart Home Personal Assistants demonstrates the potential for security frameworks to leverage empirically derived privacy norms to establish default protection levels that align with user expectations [42] .

Application across Implementation Technologies:

The framework must adapt to different levels of system sophistication whilst maintaining core protective principles:

Controlled implementations (WoZ, scripted systems): Focus on consistent application of predefined protective mechanisms and careful evaluation of persona design choices, as demonstrated in Chapter 4's implementation of standardised response protocols.

Interactive narrative platforms (Twine-like systems): Implement branching safeguards that activate appropriate protections based on scenario context and user responses, as employed in Chapter 5's domain-specific scenario design.

Dynamic LLM-based systems: Deploy real-time monitoring of manipulation patterns with adaptive interventions that respond to detected high-pressure techniques, addressing the unique vulnerabilities created by systems capable of generating contextually relevant responses in real-time [192, 193].

Alignment with Security Standards: The framework directly supports several principles in the AI Cyber Security Code of Practice [39]:

Principle 1 (Raise awareness of AI security threats and risks) through documentation of specific psychological vulnerabilities that must be considered in comprehensive AI security training. The Implementation Guide emphasises tailoring AI security training to specific roles and responsibilities, which this framework supports by demonstrating domain-specific vulnerability patterns.

Principle 4 (Enable human responsibility for AI systems) by identifying how psychological manipulation compromises human judgement. The framework provides evidence that security controls must go beyond simple awareness training, as user awareness does not necessarily translate to effective resistance.

Principle 10 (Communication and processes associated with End-users and Affected Entities) through context-sensitive protection strategies that account for domain-specific vulnerability patterns. The Implementation Guide emphasises the importance of accessible guidance for users, and this framework demonstrates how guidance should be tailored to address domain-specific vulnerabilities.

Principle 12 (Monitor your system's behaviour) by extending anomaly detection to include psychological manipulation patterns. The temporal patterns of user compliance identified in Chapter 6 provide specific indicators that could be monitored to detect manipulation attempts in real-time.

The framework also aligns with broader regulatory initiatives, including China's generative AI regulations [204], the EU AI Act [203], and the US "Blueprint for an AI Bill of Rights" [205], all of which emphasise transparent decision-making, algorithmic fairness, and robust oversight. However, as noted in recent analysis [207], these frameworks remain insufficiently detailed to explicitly address psychological manipulation, highlighting the contribution of this empirically grounded framework.

7.3.8 Framework Limitations and Future Development

Whilst this framework addresses the recognition-behaviour gap and psychological vulnerabilities identified in the empirical studies, several limitations warrant acknowledgement:

Scope Constraints:

The framework focuses specifically on psychological manipulation in conversational AI contexts and does not comprehensively address all ethical concerns in AI systems (such as algorithmic bias, environmental impacts, or labour displacement). The framework complements rather than replaces comprehensive AI ethics approaches [208, 206].

Implementation Challenges:

Operationalising the framework requires developing specific metrics, monitoring tools, and intervention mechanisms that remain subjects for future research and development. The transition from principle-based ethical frameworks to practical implementation guidance represents a significant challenge [213], though recent work on responsible AI engineering provides promising methodologies [219].

Regulatory Alignment:

The framework provides guidance for ethical design but cannot substitute for comprehensive regulatory frameworks with enforcement mechanisms. The concerning disconnect between recognised ethical imperatives and actual governance practices [207] highlights the need for regulatory frameworks that extend beyond corporate self-regulation.

Evolving Landscape:

As LLM capabilities continue advancing, new manipulation techniques may emerge that require framework extensions or modifications. The pervasive problems exhibited by LLMs, including biases, privacy concerns, copyright infringements, and lack of explainability [205], suggest that psychological manipulation represents one concern within a

broader landscape of AI risks. The lack of explainability particularly exacerbates the recognition-behaviour gap, users cannot trace the origins of persuasive content, making critical assessment significantly more difficult [205].

Mental Health Impacts:

The framework addresses manipulation in general conversational AI contexts but may require specialisation for particularly sensitive applications such as mental health support [213], where documented cases of psychological harm resulting from manipulative interactions demand heightened protections [205].

Future work should focus on developing concrete implementation methodologies for each dimension, creating standardised assessment tools for measuring psychological vulnerability, and conducting empirical studies evaluating the framework's effectiveness in protecting users whilst maintaining beneficial AI functionality. The framework should also be extended to address the unique ethical challenges introduced by generative AI's unprecedented capacity to mimic human interaction patterns [215, 205].

7.4 Limitations

While Chapters 4, 5 and 6 each address study, specific limitations and their mitigation strategies, this section synthesises the overarching limitations affecting the thesis as a whole. These limitations shape the interpretation of findings across all three empirical studies and inform the boundaries of the contributions made by this research. By acknowledging these constraints, I provide context for understanding both what the thesis demonstrates and what remains for future investigation.

7.4.1 Scope Limitations

My research focused specifically on three adversarial techniques (Need & Greed, Time, and Social Compliance) derived from the Stajano and Wilson framework of seven principles [1]. As I discussed in chapter 3, the remaining four principles (Distraction, Herd, Dishonesty, and Kindness) were not easily operationalised within conversational sys-

tems. While enabling more investigation, this selective focus means that other potentially important techniques have not been fully explored in the conversational system context.

Furthermore, I emphasised user-facing psychological vulnerabilities rather than the comprehensive security assessment described in the AI Cyber Security Code of Practice and its Implementation Guide. While psychological manipulation represents a significant security concern, it is one component of a broader security landscape that includes technical vulnerabilities, supply chain risks, and governance considerations. This limitation reflects the necessary limits of a single thesis but suggests opportunities for integrative research.

7.4.2 Generalisability Across Populations and Contexts

The participant samples across all three studies shared similar characteristics: recruitment through academic networks (Chapter 4) or Prolific Academic (Chapters 5 and 6), predominantly higher education levels, and English language fluency. Whilst this consistency enables meaningful comparison across studies, it limits generalisability to populations with different educational backgrounds, cultural contexts, or technological familiarity. Vulnerability patterns may manifest differently amongst users with varied digital literacy, age demographics, or cultural norms regarding authority and compliance.

Furthermore, the thesis examined conversational systems within specific domain contexts (healthcare, financial management, online shopping, home security, and neutral scenarios). Whilst these domains represent important application areas, they do not exhaust the range of contexts where conversational systems operate. Educational, legal, governmental, and employment contexts, amongst others, may exhibit distinct vulnerability patterns not captured in this research.

7.4.3 Cross-Study Methodological Considerations

Certain methodological patterns recur across all three studies, creating consistent constraints on the research. The within-subject experimental designs employed throughout (Chapters 4, 5 and 6) maximised statistical efficiency but introduced potential sensitisation effects as participants encountered multiple adversarial techniques within single sessions. Whilst Latin Square randomisation and counterbalancing procedures mitigated these concerns, the fundamental tension between ecological validity and experimental control persisted across all studies.

Similarly, all three studies relied substantially on self-report measures to assess attitudes and perceptions. Whilst Chapter 6 incorporated behavioural indicators through pattern matching, the predominant reliance on questionnaire data throughout the thesis means that findings reflect participants' conscious interpretations of their experiences rather than purely objective behavioural measures. This represents a consistent limitation across the research programme.

7.4.4 Methodological Limitations

The experimental settings I used in all three studies, while carefully designed to balance control and ecological validity, differ from real-world deployment contexts in important ways. The participants were aware that they were participating in the research, which could have increased their vigilance and attentiveness to security concerns. Furthermore, the stakes involved in experimental tasks, though carefully scaffolded to feel authentic, may not fully replicate the consequences of real-world decisions in contexts such as healthcare or financial management.

The technical implementation decisions in the LLM study represent another limitation that deserves consideration. The specific prompt engineering approaches used to operationalise adversarial techniques, particularly the detailed system messages and role instructions described in chapter 6 represent one of many possible implementations. Different prompt designs, system message structures, or parameter settings (such as temperature values controlling response randomness) might yield different vulnerability

patterns. Furthermore, our implementation relied on Llama 3.3-70b, and vulnerability patterns might manifest differently across other LLM architectures with different training methodologies or alignment approaches. The pattern recognition system, while sophisticated, also prioritised certain linguistic patterns over others, potentially overlooking more subtle expressions of compliance or resistance. These implementation-specific limitations highlight the need for cross-model testing approaches in future research to determine which vulnerability patterns are universal across implementations and which are model-specific.

The demographic data of the sample represents another limitation. Participants were recruited primarily through academic networks and Prolific, potentially resulting in a sample with higher education levels and greater technological familiarity than the general population. This may affect the generalisability of findings, particularly as vulnerability to psychological manipulation may vary with factors such as digital literacy and prior experience with conversational systems.

Temporal constraints represent a final methodological limitation, as my research captured responses to adversarial techniques within relatively short interaction periods. While chapter 6 included an analysis of how compliance patterns evolved during interactions, these observations covered a limited period of time. Users may develop different response patterns after extended exposure to conversational systems, potentially becoming either more resistant to manipulation through experience or more susceptible through habituation and trust development.

7.4.5 Integrating Limitations Across the Research Programme

These limitations, whilst substantial, were addressed through deliberate design choices that prioritised internal validity and systematic progression of methodological sophistication. The chapter-specific mitigation strategies detailed in Sections 4.4.6, 5.5 and 6.5 demonstrate how each study balanced constraints against research objectives. Importantly, limitations identified in earlier studies directly informed refinements in subsequent investigations, creating a methodological progression that addressed weaknesses whilst building on strengths.

The recognition of these limitations establishes boundaries for the contributions made by this thesis whilst simultaneously identifying promising directions for future research. Section 7.7 elaborates how addressing these constraints can extend and deepen understanding of psychological vulnerabilities in conversational systems.

7.5 Discussion

The findings from my research offer significant insights into the psychological security of conversational systems, with implications that extend beyond the specific techniques examined in this thesis. This section discusses broader theoretical implications, future directions for security frameworks, and considerations for the evolution of AI systems.

7.5.1 Implications for AI Security Frameworks

My research indicates that future AI security frameworks will need to evolve beyond their current focus on technical vulnerabilities to effectively address psychological security concerns. While existing frameworks like the AI Cyber Security Code of Practice [39] acknowledge human factors, they generally treat psychological manipulation as a secondary concern rather than a primary security vector.

Current security approaches have made significant advances in addressing technical vulnerabilities through sophisticated threat detection and response mechanisms. Recent empirical research with 22 experienced threat hunters [220] exemplifies this progress, revealing how cybersecurity professionals employ sophisticated techniques to identify and respond to technical threats through systematic analysis of indicators of compromise (IoCs) and attack patterns. This work demonstrates the maturity of technical threat hunting as a security practice, establishing robust methodologies for detecting and mitigating technically-orientated security breaches.

However, whilst technical security approaches have advanced considerably, psychological manipulation represents a distinct security dimension that operates through different mechanisms. Unlike technical exploits that can be detected through system logs and network traffic analysis, psychological vulnerabilities exploit human decision-

making processes, trust, and cognitive biases during interaction. This distinction highlights the need for security frameworks that address both technical and psychological dimensions in parallel, particularly as conversational AI systems create new interaction surfaces where psychological manipulation can occur alongside traditional technical threats.

The present research complements existing technical security approaches by systematically examining psychological vulnerabilities in conversational systems. Rather than replacing or superseding technical security measures, this work extends the security discourse to encompass the psychological dimensions of human-AI interaction, recognising that comprehensive security requires addressing both technical exploits and psychological manipulation techniques.

Developing unified assessment approaches that combine technical and psychological security could benefit from empirical studies of user privacy expectations, similar to the work conducted by Abdi et al. [42] on privacy norms for Smart Home Personal Assistants. Their large-scale survey of 1,738 participants identified clear patterns regarding the acceptability of information flows within complex technological ecosystems, emphasising recipient relevance as a critical determinant of acceptability. This research highlights the potential for future security frameworks to leverage empirically derived privacy norms to establish default protection levels that align closely with user expectations, especially when assessing the ethical boundaries of persuasive techniques in conversational systems.

The systematic characterisation of psychological vulnerabilities presented in this thesis suggests several directions for enhancing security frameworks. First, frameworks should incorporate dynamic vulnerability assessments that account for how user responses evolve over extended interactions, rather than focusing solely on point-in-time evaluations. The temporal patterns observed in chapter 6 demonstrate that vulnerability is not static but changes as users become more familiar with systems and techniques.

Second, security frameworks should adopt domain-specific evaluation criteria that account for the significant contextual variations observed across healthcare, financial, and other domains. The substantial differences in vulnerability patterns across these

contexts suggest that one-size-fits-all security standards may be insufficient. Instead, frameworks might incorporate domain-specific vulnerability thresholds and mitigation requirements, particularly for applications in sensitive areas such as healthcare.

Third, future frameworks will need to address the recognition-behaviour gap by incorporating structural safeguards rather than relying exclusively on awareness training. The finding that users can recognise manipulation attempts without demonstrating proportional resistance suggests that education alone is insufficient protection. Frameworks might require system-level interventions such as mandatory cooling-off periods for sensitive decisions or structural friction points that engage deliberative thinking when high-risk information is requested.

As conversational systems become more integrated into critical infrastructure and services, security frameworks must evolve to address both the technical and psychological dimensions of security. The AI Cyber Security Code of Practice represents an important step in this direction, but future iterations will need to incorporate a more detailed understanding of how psychological vulnerabilities manifest across different contexts and implementation technologies.

7.5.2 Cross-Domain Vulnerability Patterns

The significant variation in vulnerability patterns across domains has important theoretical implications for understanding human-AI interaction. Traditional models of human-computer interaction often treat the domain context as a secondary factor, focusing instead on universal principles of usability and user experience. However, my findings suggest that domain context fundamentally shapes how users interpret and respond to system behaviours, including potential manipulation attempts.

These domain effects may be explained by several theoretical mechanisms. First, contextual framing likely activates different mental models or expectations about appropriate information disclosure, shaped by prior experiences and societal norms. For example, healthcare contexts typically operate under strong privacy norms protected by regulations, whereas online shopping scenarios involve more routine data sharing. These pre-existing mental models likely influence how users evaluate information re-

quests in different domains.

Second, risk evaluation appears to be domain-specific rather than generalised. Participants demonstrated different thresholds for detecting and resisting manipulation attempts across domains, suggesting that risk perception is calibrated differently based on contextual factors. This challenges standardised security approaches such as the Common Vulnerability Scoring System (CVSS) [221], which applies uniform vulnerability metrics across different contexts, regardless of domain-specific factors that might influence user risk perception and response.

Third, the thematic analysis in chapter 5 revealed that "Agent characteristics" which emerged as a key theme, are evaluated differently across domains. Participants attributed greater legitimacy to healthcare agents compared to online shopping agents, indicating that perceived expertise and authority are context-dependent rather than universal. This suggests that theories of trust in AI systems may need to incorporate domain-specific factors rather than treating trust as a generalised construct.

These theoretical implications extend beyond security concerns to inform a broader understanding of human-AI interaction. As AI systems increasingly specialise in different domains, designers will need to consider how domain context shapes user expectations, trust dynamics, and information-sharing behaviours. Security measures that fail to account for these domain-specific patterns may be ineffective or, worse, disruptive to legitimate interaction patterns.

7.5.3 Evolution of Vulnerability with Implementation Sophistication

As conversational systems continue to advance in capability, new vulnerability patterns are likely to emerge that extend beyond those observed in this thesis. The progression from a Wizard of Oz methodology to LLM-based implementation revealed how vulnerability patterns shift as systems become more sophisticated and autonomous. The increasingly naturalistic interactions enabled by advanced LLMs will likely blur the boundaries between human and machine communication, potentially magnifying social response patterns that bypass critical evaluation. At the same time, more sophisticated systems may enable new forms of user protection by analysing interaction patterns in

real time and detecting manipulation attempts early.

The recognition-behaviour gap observed in chapter 6 may become more pronounced as systems become more capable of generating human-like responses. This has significant implications for AI alignment research, which often assumes that transparent behaviour will enable effective human oversight. If users can recognise problematic system behaviours but still comply with them, transparency alone may be insufficient to ensure safe and beneficial AI.

As generative AI capabilities continue to advance, it is important to consider how the psychological vulnerabilities identified in this thesis might evolve. Systems that can generate increasingly personalised content may create more effective versions of the Need & Greed technique by targeting individual preferences and motivations. Similarly, systems with access to broader contextual information may implement more sophisticated Social Compliance techniques by leveraging social network effects and group dynamics.

7.6 Conclusion

This thesis has investigated the operationalisation of adversarial techniques in conversational systems, examining how psychological principles derived from scam tactics can be implemented to influence user attitudes and behaviours. Through a progressive series of three user studies, I have demonstrated that Need & Greed, Time, and Social Compliance techniques significantly impact user comfort levels and willingness to disclose information, with effects varying substantially by context and implementation approach.

My research makes important contributions to both theoretical understanding and practical application in conversational system security. Theoretically, it extends Stajano and Wilson [1]'s framework of scam principles to the domain of conversational AI, providing empirical validation of how these principles manifest in human-AI interaction. Methodologically, it progresses from controlled to dynamic assessment techniques, offering a framework for future research on psychological security in AI systems.

The findings reveal complex vulnerability patterns that vary by technique, context, and implementation. Social compliance techniques achieve higher overall compliance rates than Need & Greed approaches but operate through more subtle, gradual influence mechanisms. Context significantly moderates the effectiveness of techniques, with healthcare scenarios demonstrating distinctive resistance patterns compared to commercial contexts. Implementation sophistication introduces additional complexity, with LLM-based systems creating unique vulnerability patterns, including the recognition-behaviour gap and temporal influence patterns.

Looking back at my thesis statement (section 1.3), this research has demonstrated that adversarial techniques derived from scam principles can be systematically operationalised in conversational systems, with their effectiveness varying significantly based on context, domain, and implementation technology. Understanding these contextual variations and implementation effects is crucial for identifying psychological vulnerabilities in human-AI interactions and establishing a foundation for future security measures.

By identifying and characterising these psychological vulnerabilities, this thesis contributes to a more comprehensive understanding of security vulnerabilities in conversational systems. Rather than providing design solutions, it establishes the problem space that must be addressed, creating a foundation for future work on security approaches that can effectively protect users from the psychological manipulation techniques identified in this research. Specifically, future security measures will need to address the recognition-behaviour gap, where awareness fails to trigger proportional defence; the temporal dynamics of different techniques, which require monitoring over extended interactions rather than point-in-time assessments; and domain-specific vulnerability patterns that necessitate context-aware protection strategies. By empirically documenting these vulnerability patterns across techniques, contexts, and implementations, this research lays the essential groundwork for developing theoretically sound, empirically validated security approaches for conversational systems.

In conclusion, this thesis demonstrates that conversational systems are not merely vulnerable to technical exploits but critically exposed to psychological manipulation

techniques that directly influence human decision-making. My research emphasises that psychological vulnerabilities, such as susceptibility to social authority, time pressure, and incentives, must be considered central, rather than peripheral, in the design and evaluation of secure conversational systems. As these systems increasingly integrate into essential services like healthcare, finance, and personal security, overlooking psychological vulnerabilities could severely compromise user autonomy and privacy. Thus, the future of AI security necessitates a fundamental shift from technical defence alone towards integrated frameworks that combine psychological insights, empirical user-centred research, and context-aware protections. Only by systematically addressing both the technical and psychological dimensions can we build truly trustworthy, ethically robust, and resilient conversational systems.

7.7 Future Work

The findings and limitations of my research suggest several promising directions for future research, ranging from direct extensions of the current work to entirely new approaches that build on its theoretical and methodological foundations.

7.7.1 Direct Extension

Refined Application of Selected Principles

As I established in chapter 3, the remaining four principles of the Stajano and Wilson framework (distraction, herd, dishonesty, and kindness) were identified as not easily operationalised within conversational systems. However, future work could explore innovative approaches to adapting these principles for text-based or multimodal conversational contexts. For instance, the Distraction Principle, which relies heavily on physical misdirection in traditional scams, might be reconceptualised for conversational systems through information overload or topic-shifting techniques that overwhelm users' cognitive capacities.

Longitudinal studies represent another valuable extension that examines how users

adapt to adversarial techniques over time. Based on the temporal patterns observed in chapter 6, such research could investigate whether the immediate effectiveness of Need & Greed techniques diminishes with repeated exposure and whether the gradual influence of Social Compliance continues to build over extended interactions. This temporal dimension is particularly important as conversational systems become more integrated into daily routines and long-term relationships between users and systems develop.

Finally, extending domain coverage to additional contexts, such as education, entertainment, and workplace assistance, would further enhance our understanding of context-specific vulnerability patterns. Cross-cultural investigations would be especially valuable, examining how cultural factors influence susceptibility to different adversarial techniques and informing globally appropriate security measures.

Security Standard Alignment

Future work could focus on developing testing methodologies specifically for psychological security aspects of the AI Cyber Security Code of Practice and its Implementation Guide. As the Implementation Guide notes, organisations should "evaluate threats and manage risks" to AI systems, but currently lack specific guidance on assessing psychological vulnerabilities. My research could inform the development of standardised assessment tools to measure human vulnerability in AI systems, potentially incorporating metrics for resistance to manipulation alongside technical security measures. Such tools would help organisations implement comprehensive security evaluations that address both technical and psychological aspects of system security.

Another promising direction is to create empirical foundations for future iterations of security standards. By systematically investigating the effectiveness of different countermeasures against psychological manipulation, this research could inform evidence-based security requirements that balance protection, usability, and implementation feasibility. This aligns with the Implementation Guide's emphasis on continually improving security measures in response to emerging threats and vulnerabilities.

7.7.2 New Approach

Cross-Model Adversarial Testing Framework

While human studies are critical for understanding psychological vulnerabilities, they are also resource-intensive. The proposed Cross-Model adversarial testing framework (built on the foundations established by the Flask-based experimental framework in chapter 6) serves as an innovative supplementary approach to rapidly screen for conversational manipulations across various LLMs. However, a significant limitation is that LLMs lack genuine psychological processes; thus, findings from cross-model testing may not reliably transfer directly to human vulnerabilities. The main value of this approach is its ability to swiftly identify prompts or conversational strategies likely to be effective at a system level, offering useful preliminary insights. These initial findings can guide more resource-intensive human-based studies, ultimately refining experimental hypotheses and enabling targeted empirical investigations of psychological vulnerabilities. Furthermore, this framework may be valuable for identifying inherent vulnerabilities within conversational systems, potentially informing system-level defences alongside human-centred interventions.

In this framework, one model (e.g. Llama) would be configured with adversarial techniques (Need & Greed, Time, or Social Compliance), while another model (e.g. GPT[1]) from a different architecture would act as the target, simulating various user vulnerability profiles. This cross-model testing offers several advantages over traditional experimental approaches.

1. Architectural diversity: Testing across different foundation models would reveal whether susceptibility to adversarial techniques varies withfor identifying inherent vulnerabilities within conversational systems model architecture, training methodologies, or alignment approaches.

2. Systematic evaluation: The framework would enable thousands of controlled interactions to be executed efficiently, enabling statistical analysis of the effective-

[1]GPT-4: https://openai.com/gpt-4

ness of the technique across multiple domains and vulnerability profiles.

3. Technique optimisation: By implementing focused technique testing, the research could systematically refine the understanding of how specific adversarial approaches function in various contexts.

The implementation would require developing a unified interface layer to harmonise the different API structures, response formats, and parameter configurations between model providers. Additionally, an orchestration system would be needed to manage the conversation flow, record interactions, and calculate effectiveness metrics based on the patterns established in this thesis.

Given the technical challenges, a phased implementation approach would be advisable, starting with a single technique (e.g., Need & Greed) across multiple domains before expanding to other techniques. This would allow for focused analysis of technique-specific patterns while managing the computational and financial resources required for large-scale testing. The results from such cross-model testing could have significant implications for AI security standards. They could potentially reveal which model architectures offer inherent resistance to psychological manipulation and which architectural features contribute to vulnerability. This information could inform both defensive strategies for conversational system design and evaluation criteria for security-focused model selection, supporting the implementation of the Code of Practice's security-by-design principles.

Addressing the Recognition-Behaviour Gap

The recognition-behaviour gap identified in chapter 6 represents a promising area for future research. This phenomenon, in which users identify manipulation attempts without demonstrating corresponding resistance behaviours, suggests that traditional security awareness approaches may be ineffective against psychological manipulation. Future work could explore why this gap exists and develop interventions that effectively bridge recognition and protective behaviour.

This research could investigate cognitive, emotional, and contextual factors contributing to the gap, including cognitive load, social pressure, perceived benefits, and risk assessment processes. By understanding the psychological mechanisms that maintain this gap, more effective protection strategies could be developed that address root causes rather than symptoms.

Intervention studies could test different approaches to closing the gap, ranging from enhanced awareness training to structural safeguards that reduce the cognitive burden of security decisions. These might include just-in-time security nudges that align with users' recognition of manipulation attempts or interaction design patterns that automatically increase friction when manipulation is detected, providing users with additional time and support for protective decision-making.

Integrated Security Frameworks

Developing unified assessment approaches that combine technical and psychological security represents another innovative direction for future research. Such frameworks would recognise the interdependence of these security domains, acknowledging that technical vulnerabilities can enable psychological manipulation and vice versa. This integrated approach aligns with the comprehensive security vision articulated in the AI Cyber Security Code of Practice, but requires new methodologies that span traditionally separate domains. The Implementation Guide emphasises the importance of conducting threat modelling that includes AI-specific attacks, but currently focuses primarily on technical vulnerabilities. My research suggests that psychological vulnerabilities should be modelled and addressed in a similar way. Future work could develop integrated assessment frameworks that consider both technical and psychological attack vectors, providing a more comprehensive security evaluation. Creating design patterns that inherently resist manipulation techniques offers another promising approach. Rather than adding security measures to existing designs, this would involve developing fundamental interaction paradigms that minimise vulnerability by design. Examples might include interaction patterns that naturally prompt critical evaluation or make manipulation attempts more visible to users. This aligns with the

Implementation Guide's emphasis on security-by-design principles.

Defensive Design Patterns

Future research could develop and evaluate specific defensive design patterns to counter the adversarial techniques identified in this thesis. For Need & Greed, this might include transparent benefit structures that highlight potential risks alongside rewards. For Time pressure, it could involve design patterns that encourage deliberation even in seemingly urgent situations. For Social Compliance, it might include mechanisms to verify authority, helping users distinguish legitimate from manipulative authority claims.

These defensive patterns would support the implementation of Principle 2 of the Code (Design your AI system for security, functionality, and performance) by providing practical design approaches that mitigate psychological vulnerabilities. The Implementation Guide emphasises that systems should be designed to withstand adversarial attacks, and these patterns would directly address this requirement in the face of psychological manipulation attempts.

Training and awareness approaches represent another valuable area for investigation. Research could examine the effectiveness of different educational interventions in building user resistance to manipulation, potentially incorporating simulated adversarial interactions as learning experiences. The goal would be to develop evidence-based training approaches that enhance security without creating an unnecessary burden or anxiety for users. This aligns with the Implementation Guide's emphasis on security training, extending it to include detection and resistance to psychological manipulation, as outlined.

Security Governance Approaches

Translating research findings into specific policy recommendations represents an important direction for future work. This could involve developing detailed guidelines for regulators and standards bodies that outline how psychological security considerations

should be incorporated into AI governance frameworks. Such research would bridge academic insights with practical governance needs, helping to ensure that emerging regulations address the full spectrum of security concerns. Developing certification criteria for psychologically secure conversational systems offers another direction of governance-orientated research. This would involve developing assessment methodologies to verify compliance with security standards in a reliable and transparent manner. Such certification approaches would help users and organisations make informed decisions about which conversational systems to trust with sensitive interactions, supporting the implementation of the Code of Practice's accountability and governance principles.

7.8 Summary

This chapter integrated findings from the three user studies presented in this thesis to address the research questions and highlight contributions to the field. Beginning with a synthesis of how adversarial techniques influence user attitudes and behaviours across different contexts and implementation technologies, I then explored contributions, limitations, broader implications, and future directions.

First, section 7.1 directly addressed the overarching research question by synthesising findings from chapter 4 to chapter 6, demonstrating how Need & Greed, Time, and Social Compliance techniques influence user attitudes and behaviours. This section showed how these effects vary by technique type, domain context, implementation technology, and user factors, creating a comprehensive answer to the central question posed in the thesis statement.

Building on this synthesis, section 7.2 outlined three primary contributions to the field: empirical validation of psychological vulnerabilities, alignment with security frameworks, and methodological progression. This section explained how each study built upon the previous one to create a comprehensive understanding of conversational system security, highlighting key findings, such as the recognition-behaviour gap, in which users recognised manipulation attempts without demonstrating proportional re-

sistance.

section 7.3 presented a conceptual framework for addressing psychological vulnerabilities in conversational AI systems. This framework synthesises empirical findings from the three user studies into five integrated dimensions: Autonomy Preservation by Design, Structural Safeguards Beyond Awareness, Context-Sensitive Ethics, Persona Consistency and Transparency, and Continuous Vulnerability Monitoring. The framework directly addresses the recognition-behaviour gap identified in chapter 6, proposing systematic approaches to protect user autonomy when awareness alone proves insufficient. By aligning with principles in the AI Cyber Security Code of Practice and broader regulatory initiatives, the framework demonstrates how empirical findings can inform practical security approaches whilst acknowledging limitations in scope, implementation challenges, and the need for ongoing development as LLM capabilities continue advancing.

section 7.4 acknowledged research limitations, including scope constraints, methodological considerations, and sample characteristics that affect the generalisability of the findings. The discussion in section 7.5 explored the broader implications of the research, examining how the identified vulnerabilities relate to AI security frameworks, how they vary across different domains, and how they evolve with implementation sophistication. This section emphasised how these findings establish an empirical foundation for future security research rather than providing direct design solutions.

Finally, section 7.6 concluded by reaffirming the thesis statement and summarising the key contributions, while section 7.7 outlined promising directions for future research. These include refined application of selected principles, alignment of security standards, cross-model adversarial testing, addressing the recognition-behaviour gap, and developing integrated security frameworks. This chapter established how this research has advanced our understanding of psychological vulnerabilities in conversational systems and created a foundation for future work in this important area.

Bibliography

[1] F. Stajano and P. Wilson, "Understanding scam victims: seven principles for systems security," *Communications of the ACM*, vol. 54, no. 3, pp. 70–75, Mar. 2011. [Online]. Available: https://doi.org/10.1145/1897852.1897872

[2] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera, "Conversational agents in healthcare: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, Sep. 2018. [Online]. Available: https://doi.org/10.1093/jamia/ocy072

[3] L. Balcombe and D. De Leo, "Human-Computer Interaction in Digital Mental Health," *Informatics*, vol. 9, no. 1, p. 14, Mar. 2022, number: 1 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2227-9709/9/1/14

[4] J. Balakrishnan and Y. K. Dwivedi, "Conversational commerce: entering the next stage of AI-powered digital assistants," *Annals of Operations Research*, Apr. 2021. [Online]. Available: https://doi.org/10.1007/s10479-021-04049-5

[5] C. Rodrigues, A. Reis, R. Pereira, P. Martins, J. Sousa, and T. Pinto, "A Review of Conversational Agents in Education," in *Technology and Innovation in Learning, Teaching and Education*, ser. Communications in Computer and Information Science, A. Reis, J. Barroso, P. Martins, A. Jimoyiannis, R. Y.-M. Huang, and R. Henriques, Eds. Cham: Springer Nature Switzerland, 2022, pp. 461–467.

[6] A. Shemshack and J. M. Spector, "A systematic literature review of personalized learning terms," *Smart Learning Environments*, vol. 7, no. 1, p. 33, Oct. 2020. [Online]. Available: https://doi.org/10.1186/s40561-020-00140-9

[7] M. Adam, M. Wessel, and A. Benlian, "AI-based chatbots in customer service and their effects on user compliance," *Electronic Markets*, vol. 31, no. 2, pp. 427–445, Jun. 2021. [Online]. Available: https://doi.org/10.1007/s12525-020-00414-7

[8] C. B. Nordheim, A. Følstad, and C. A. Bjørkli, "An Initial Model of Trust in Chatbots for Customer Service—Findings from a Questionnaire Study," *Interacting with Computers*, vol. 31, no. 3, pp. 317–335, May 2019. [Online]. Available: https://academic.oup.com/iwc/article/31/3/317/5553672

[9] E. Ruane and V. Nallur, ""EHLO WORLD" – Checking If Your Conversational AI Knows Right from Wrong," Jun. 2020, arXiv:2006.10437 [cs]. [Online]. Available: http://arxiv.org/abs/2006.10437

[10] S. Schöbel, A. Schmitt, D. Benner, M. Saqr, A. Janson, and J. M. Leimeister, "Charting the Evolution and Future of Conversational Agents: A Research Agenda Along Five Waves and New Frontiers," *Information Systems Frontiers*, vol. 26, no. 2, pp. 729–754, Apr. 2024. [Online]. Available: https://doi.org/10.1007/s10796-023-10375-9

[11] W. Liu, K. Xu, and M. Z. Yao, ""Can you tell me about yourself?" The impacts of chatbot names and communication contexts on users' willingness to self-disclose information in human-machine conversations," *Communication Research Reports*, vol. 40, no. 3, pp. 122–133, May 2023, publisher: Routledge _eprint: https://doi.org/10.1080/08824096.2023.2212899. [Online]. Available: https://doi.org/10.1080/08824096.2023.2212899

[12] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot," *Comput. Linguist.*, vol. 46, no. 1, pp. 53–93, Mar. 2020. [Online]. Available: https://dl.acm.org/doi/10.1162/coli_a_00368

[13] M. Hasal, J. Nowaková, K. Ahmed Saghair, H. Abdulla, V. Snášel, and L. Ogiela, "Chatbots: Security, privacy, data protection, and social aspects,"

Concurrency and Computation: Practice and Experience, vol. 33, no. 19, Oct. 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/cpe.6426

[14] M. Larson, N. Oostdijk, and F. Zuiderveen Borgesius, "Not Directly Stated, Not Explicitly Stored: Conversational Agents and the Privacy Threat of Implicit Information," in *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '21. New York, NY, USA: Association for Computing Machinery, Jun. 2021, pp. 388–391. [Online]. Available: https://doi.org/10.1145/3450614.3463601

[15] B. Marchegiani, "Anthropomorphism, false beliefs, and conversational ais: How chatbots undermine users' autonomy," *Journal of Applied Philosophy*, 2025.

[16] N. Mireshghallah, M. Antoniak, Y. More, Y. Choi, and G. Farnadi, "Trust no bot: Discovering personal disclosures in human-llm conversations in the wild," *arXiv preprint arXiv:2407.11438*, 2024.

[17] S. Tran, H. Lu, I. Slaughter, B. Herman, A. Dangol, Y. Fu, L. Chen, B. Gebreyohannes, B. Howe, A. Hiniker *et al.*, "Understanding privacy norms around llm-based chatbots: A contextual integrity perspective," *arXiv preprint arXiv:2508.06760*, 2025.

[18] F. Salvi, M. H. Ribeiro, R. Gallotti, and R. West, "On the conversational persuasiveness of large language models: A randomized controlled trial," *arXiv preprint arXiv:2403.14380*, 2024.

[19] A. Rogiers, S. Noels, M. Buyl, and T. De Bie, "Persuasion with large language models: a survey," *arXiv preprint arXiv:2411.06837*, 2024.

[20] A. Mollazehi, I. Abuelezz, M. Barhamgi, K. M. Khan, and R. Ali, "Do cialdini's persuasion principles still influence trust and risk-taking when social engineering is knowingly possible?" in *International Conference on Research Challenges in Information Science*. Springer, 2024, pp. 273–288.

[21] M. R. Rahman, S. K. Basak, R. M. Hezaveh, and L. Williams, "Attackers reveal their arsenal: An investigation of adversarial techniques in cti reports," *arXiv preprint arXiv:2401.01865*, 2024.

[22] S. Quach, P. Thaichon, K. D. Martin, S. Weaven, and R. W. Palmatier, "Digital technologies: tensions in privacy and data," *Journal of the Academy of Marketing Science*, vol. 50, no. 6, pp. 1299–1323, Nov. 2022. [Online]. Available: https://doi.org/10.1007/s11747-022-00845-y

[23] E. Aïmeur, N. Díaz Ferreyra, and H. Hage, "Manipulation and Malicious Personalization: Exploring the Self-Disclosure Biases Exploited by Deceptive Attackers on Social Media," *Frontiers in Artificial Intelligence*, vol. 2, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2019.00026

[24] C. Liu and Y. Wei, "The impacts of time constraint on users' search strategy during search process," *Proceedings of the Association for Information Science and Technology*, vol. 53, no. 1, pp. 1–9, 2016, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2016.14505301051. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2016.14505301051

[25] C. M. Wu, E. Schulz, T. J. Pleskac, and M. Speekenbrink, "Time pressure changes how people explore and respond to uncertainty," *Scientific Reports*, vol. 12, no. 1, p. 4122, Mar. 2022, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-022-07901-1

[26] S. P. Saunderson and G. Nejat, "Persuasive robots should avoid authority: The effects of formal and real authority on persuasion in human-robot interaction," *Science Robotics*, vol. 6, no. 58, p. eabd5186, Sep. 2021, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.abd5186

[27] A. S. Ghazali, J. Ham, E. Barakova, and P. Markopoulos, "Assessing the effect of persuasive robots interactive social cues on users' psychological reactance, liking, trusting beliefs and compliance," *Advanced Robotics*, vol. 33, no. 7-8, pp. 325–337, Apr. 2019, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01691864.2019.1589570. [Online]. Available: https://doi.org/10.1080/01691864.2019.1589570

[28] R. Cialdini, "The Science of Persuasion," *Scientific American*, Jan. 2001. [Online]. Available: https://www.academia.edu/30721147/The_Science_of_Persuasion

[29] Y. Trope and N. Liberman, "Temporal construal," *Psychological Review*, vol. 110, no. 3, pp. 403–421, 2003, place: US Publisher: American Psychological Association.

[30] G. Murtarelli, A. Gregory, and S. Romenti, "A conversation-based perspective for shaping ethical human–machine interactions: The particular challenge of chatbots," *Journal of Business Research*, vol. 129, pp. 927–935, May 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296320305944

[31] J. Kwesi, J. Cao, R. Manchanda, and P. Emami-Naeini, "Exploring user security and privacy attitudes and concerns toward the use of general-purpose llm chatbots for mental health," *arXiv preprint arXiv:2507.10695*, 2025.

[32] V. Bakir, A. Laffer, A. McStay, D. Miranda, and L. Urquhart, "On manipulation by emotional ai: Uk adults' views and governance implications," *Frontiers in Sociology*, vol. 9, p. 1339834, 2024.

[33] E. Gumusel, "A literature review of user privacy concerns in conversational chatbots: A social informatics approach: An annual review of information science and technology (arist) paper," *Journal of the Association for Information Science and Technology*, vol. 76, no. 1, pp. 121–154, 2025.

[34] R. Zhang, H. Li, A. Chen, Z. Liu, and Y.-C. Lee, "Ai privacy in context: A comparative study of public and institutional discourse on conversational ai privacy in the us and chinese social media," *Social Media+ Society*, vol. 10, no. 4, p. 20563051241290845, 2024.

[35] C. M. Fang, A. R. Liu, V. Danry, E. Lee, S. W. Chan, P. Pataranutaporn, P. Maes, J. Phang, M. Lampe, L. Ahmad *et al.*, "How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study," *arXiv preprint arXiv:2503.17473*, 2025.

[36] K. Moylan and K. Doherty, "Expert and interdisciplinary analysis of ai-driven chatbots for mental health support: Mixed methods study," *Journal of Medical Internet Research*, vol. 27, p. e67114, 2025.

[37] J. P. Vargheese, M. Collinson, and J. Masthoff, "Exploring Susceptibility Measures to Persuasion," in *Persuasive Technology. Designing for Future Change*, ser. Lecture Notes in Computer Science, S. B. Gram-Hansen, T. S. Jonasen, and C. Midden, Eds. Cham: Springer International Publishing, 2020, pp. 16–29.

[38] S. Thönes, S. Arnau, and E. Wascher, "Cognitions about time affect perception, behavior, and physiology – A review on effects of external clock-speed manipulations," *Consciousness and Cognition*, vol. 63, pp. 99–109, Aug. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053810018300904

[39] Department for Science, Innovation & Technology, "Code of Practice for the Cyber Security of AI," Jan. 2025. [Online]. Available: https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice/code-of-practice-for-the-cyber-security-of-ai

[40] R. J. Anderson, *Security Engineering: A Guide to Building Dependable Distributed Systems*, 2nd ed. Indianapolis, IN: Wiley, 2008.

[41] H. Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, CA: Stanford Law Books, 2010.

[42] N. Abdi, X. Zhan, K. M. Ramokapane, and J. Such, "Privacy Norms for Smart Home Personal Assistants," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–14. [Online]. Available: https://doi.org/10.1145/3411764.3445122

[43] R. Anderson, "Why information security is hard – an economic perspective," in *Proceedings of the 17th Annual Computer Security Applications Conference (ACSAC)*. New Orleans, LA, USA: IEEE Computer Society, 2001, pp. 358–365.

[44] R. W. Rogers, "A protection motivation theory of fear appeals and attitude change," *The Journal of Psychology*, vol. 91, no. 1, pp. 93–114, 1975.

[45] S. R. Boss, D. F. Galletta, P. B. Lowry, G. D. Moody, and P. Polak, "What do systems users have to fear? using fear appeals to engender threats and fear that motivate protective security behaviors1," *MIS quarterly*, vol. 39, no. 4, pp. 837–864, 2015.

[46] R. S. Laufer and M. Wolfe, "Privacy as a concept and a social issue: A multi-dimensional developmental theory," *Journal of Social Issues*, vol. 33, no. 3, pp. 22–42, 1977.

[47] T. Dinev and P. Hart, "An Extended Privacy Calculus Model for E-Commerce Transactions," *Information Systems Research*, vol. 17, no. 1, pp. 61–80, Mar. 2006, publisher: INFORMS. [Online]. Available: https://pubsonline.informs.org/doi/10.1287/isre.1060.0080

[48] G. Loewenstein, "Out of Control: Visceral Influences on Behavior," *Organizational Behavior and Human Decision Processes*, vol. 65, no. 3, pp. 272–292, Mar. 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S074959789690028X

[49] D. Kahneman, "Fast and slow thinking," *Allen Lane and Penguin Books, New York*, 2011.

[50] L. Sillence, P. Mo, P. Briggs, and P. R. Harris, "The Changing Face of Trust in Health Websites," Rochester, NY, Aug. 2011. [Online]. Available: https://papers.ssrn.com/abstract=1920317

[51] S. A. Saadati and S. M. Saadati, "The Role of Chatbots in Mental Health Interventions: User Experiences," *AI and Tech in Behavioral and Social Sciences*, vol. 1, no. 2, pp. 19–25, 2023. [Online]. Available: https://journals.kmanpub.com/index.php/aitechbesosci/article/view/1982

[52] F. M. Calisto, J. Fernandes, M. Morais, C. Santiago, J. M. Abrantes, N. Nunes, and J. C. Nascimento, "Assertiveness-based Agent Communication for a Personalized Medicine on Medical Imaging Diagnosis," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–20. [Online]. Available: https://dl.acm.org/doi/10.1145/3544548.3580682

[53] J. Parviainen and J. Rantala, "Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care," *Medicine, Health Care and Philosophy*, vol. 25, no. 1, pp. 61–71, Mar. 2022. [Online]. Available: https://doi.org/10.1007/s11019-021-10049-w

[54] R. Hasan, R. Shams, and M. Rahman, "Consumer trust and perceived risk for voice-controlled artificial intelligence: The case of Siri," *Journal of Business Research*, vol. 131, pp. 591–597, Jul. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296320308419

[55] M. M. Van Pinxteren, M. Pluymaekers, and J. G. Lemmink, "Human-like communication in conversational agents: a literature review and research agenda," *Journal of Service Management*, vol. 31, no. 2, pp. 203–225,

Jan. 2020, publisher: Emerald Publishing Limited. [Online]. Available: https://doi.org/10.1108/JOSM-06-2019-0175

[56] M. Aliannejadi, L. Azzopardi, H. Zamani, E. Kanoulas, P. Thomas, and N. Craswel, "Analysing Mixed Initiatives and Search Strategies during Conversational Search," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, Oct. 2021, pp. 16–26, arXiv:2109.05955 [cs]. [Online]. Available: http://arxiv.org/abs/2109.05955

[57] M. Bahja, R. Hammad, and G. Butt, "A User-Centric Framework for Educational Chatbots Design and Development," in *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, ser. Lecture Notes in Computer Science, C. Stephanidis, M. Kurosu, H. Degen, and L. Reinerman-Jones, Eds. Cham: Springer International Publishing, 2020, pp. 32–43.

[58] M. Dubiel, M. Halvey, L. Azzopardi, and S. Daronnat, "Interactive Evaluation of Conversational Agents: Reflections on the Impact of Search Task Design," in *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ser. ICTIR '20. New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 85–88. [Online]. Available: https://doi.org/10.1145/3409256.3409814

[59] J. Kiesel, L. Meyer, M. Potthast, and B. Stein, "Meta-Information in Conversational Search," *ACM Transactions on Information Systems*, vol. 39, no. 4, pp. 50:1–50:44, Aug. 2021. [Online]. Available: https://doi.org/10.1145/3468868

[60] O. Seminck, "Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots by Michael McTear," *Computational Linguistics*, vol. 49, no. 1, pp. 257–259, Mar. 2023. [Online]. Available: https://direct.mit.edu/coli/article/49/1/257/113849/Conversational-AI-Dialogue-Systems-Conversational

[61] M. Rostami and S. Navabinejad, "Artificial Empathy: User Experiences with Emotionally Intelligent Chatbots," *AI and Tech in Behavioral and Social Sciences*, vol. 1, no. 3, pp. 19–27, 2023. [Online]. Available: https://journals.kmanpub.com/index.php/aitechbesosci/article/view/1989

[62] J. Zhang, Y. J. Oh, P. Lange, Z. Yu, and Y. Fukuoka, "Artificial Intelligence Chatbot Behavior Change Model for Designing Artificial Intelligence Chatbots to Promote Physical Activity and a Healthy Diet: Viewpoint," *Journal of Medical Internet Research*, vol. 22, no. 9, p. e22845, Sep. 2020. [Online]. Available: https://www.jmir.org/2020/9/e22845

[63] E. J. Gerritse, F. Hasibi, and A. P. de Vries, "Bias in Conversational Search: The Double-Edged Sword of the Personalized Knowledge Graph," in *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ser. ICTIR '20. New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 133–136. [Online]. Available: https://doi.org/10.1145/3409256.3409834

[64] A. Ramesh and V. Chawla, "Chatbots in Marketing: A Literature Review Using Morphological and Co-Occurrence Analyses," *Journal of Interactive Marketing*, vol. 57, no. 3, pp. 472–496, Aug. 2022. [Online]. Available: https://journals.sagepub.com/doi/10.1177/10949968221095549

[65] V. T. Nguyen, L. T. Phong, and N. T. K. Chi, "The impact of AI chatbots on customer trust: an empirical investigation in the hotel industry," *Consumer Behavior in Tourism and Hospitality*, vol. 18, no. 3, pp. 293–305, Aug. 2023. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/CBTH-06-2022-0131/full/html

[66] M. Cascella, F. Semeraro, J. Montomoli, V. Bellini, O. Piazza, and E. Bignami, "The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives," *Journal of*

Medical Systems, vol. 48, no. 1, p. 22, Feb. 2024. [Online]. Available: https://doi.org/10.1007/s10916-024-02045-3

[67] P. Ren, Z. Chen, Z. Ren, E. Kanoulas, C. Monz, and M. De Rijke, "Conversations with Search Engines: SERP-based Conversational Response Generation," *ACM Transactions on Information Systems*, vol. 39, no. 4, pp. 1–29, Oct. 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3432726

[68] J. Ha, H. Jeon, D. Han, J. Seo, and C. Oh, "CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, May 2024, pp. 1–24. [Online]. Available: https://dl.acm.org/doi/10.1145/3613904.3642472

[69] Y. Deng, L. Liao, L. Chen, H. Wang, W. Lei, and T.-S. Chua, "Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10 602–10 621. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.711/

[70] Z. He, Z. Xie, H. Steck, D. Liang, R. Jha, N. Kallus, and J. McAuley, "Reindex-Then-Adapt: Improving Large Language Models for Conversational Recommendation," May 2024, arXiv:2405.12119 [cs]. [Online]. Available: http://arxiv.org/abs/2405.12119

[71] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A Comprehensive Overview of Large Language Models," Apr. 2024, arXiv:2307.06435 [cs]. [Online]. Available: http://arxiv.org/abs/2307.06435

[72] A. J. Thirunavukarasu, S. Mahmood, A. Malem, W. P. Foster, R. Sanghera, R. Hassan, S. Zhou, S. W. Wong, Y. L. Wong, Y. J. Chong, A. Shakeel,

Y.-H. Chang, B. K. J. Tan, N. Jain, T. F. Tan, S. Rauz, D. S. W. Ting, and D. S. J. Ting, "Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study," *PLOS Digital Health*, vol. 3, no. 4, p. e0000341, Apr. 2024, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000341

[73] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, and D. Gašević, "Practical and ethical challenges of large language models in education: A systematic scoping review," *British Journal of Educational Technology*, vol. 55, no. 1, pp. 90–112, 2024, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13370. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13370

[74] E. Ayedoun, Y. Hayashi, and K. Seta, "Adding Communicative and Affective Strategies to an Embodied Conversational Agent to Enhance Second Language Learners' Willingness to Communicate," *International Journal of Artificial Intelligence in Education*, 2018.

[75] L. Villa, "Comparative Analysis of Generic and Fine-Tuned Large Language Models for Conversational Agent Systems," *Robotics*, 2024.

[76] J. Wei, S. Kim, H. Jung, and Y.-H. Kim, "Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–35, Apr. 2024, arXiv:2301.05843 [cs]. [Online]. Available: http://arxiv.org/abs/2301.05843

[77] Z. Dong, Z. Zhou, C. Yang, J. Shao, and Y. Qiao, "Attacks, Defenses and Evaluations for LLM Conversation Safety: A Survey," Mar. 2024, arXiv:2402.09283 [cs]. [Online]. Available: http://arxiv.org/abs/2402.09283

[78] J. Liu, C. Symons, and R. R. Vatsavai, "Persona-Based Conversational AI: State of the Art and Challenges," in *2022 IEEE International Conference on*

Data Mining Workshops (ICDMW), Nov. 2022, pp. 993–1001, iSSN: 2375-9259. [Online]. Available: https://ieeexplore.ieee.org/document/10031078

[79] T. W. Bickmore, S. E. Mitchell, B. W. Jack, M. K. Paasche-Orlow, L. M. Pfeifer, and J. O'Donnell, "Response to a relational agent by hospital patients with depressive symptoms," *Interacting with Computers*, vol. 22, no. 4, pp. 289–298, Jul. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0953543809001064

[80] K. Kretzschmar, H. Tyroll, G. Pavarini, A. Manzini, and I. Singh, "Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support," *Biomedical Informatics Insights*, vol. 11, p. 1178222619829083, Jan. 2019. [Online]. Available: https://doi.org/10.1177/1178222619829083

[81] A. Ho, J. Hancock, and A. S. Miner, "Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot," *Journal of Communication*, vol. 68, no. 4, pp. 712–733, Aug. 2018. [Online]. Available: https://academic.oup.com/joc/article/68/4/712/5025583

[82] S. Peter, K. Riemer, and J. D. West, "The benefits and dangers of anthropomorphic conversational agents," *Proceedings of the National Academy of Sciences*, vol. 122, no. 22, p. e2415898122, 2025.

[83] K. Saffarizadeh, M. Keil, M. Boodraj, and T. Alashoor, ""my name is alexa. what's your name?" the impact of reciprocal self-disclosure on post-interaction trust in conversational agents," *Journal of the Association for Information Systems*, vol. 25, no. 3, pp. 528–568, 2024.

[84] J. Biro, C. Linder, D. Neyens *et al.*, "The effects of a health care chatbot's complexity and persona on user trust, perceived usability, and effectiveness: mixed methods study," *JMIR Human Factors*, vol. 10, no. 1, p. e41017, 2023.

[85] G. Park, J. Chung, and S. Lee, "Effect of AI chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: a serial mediation model," *Current Psychology*, vol. 42, no. 32, pp. 28 663–28 673, Nov. 2023. [Online]. Available: https://doi.org/10.1007/s12144-022-03932-z

[86] T. W. Bickmore, D. Utami, R. Matsuyama, and M. K. Paasche-Orlow, "Improving Access to Online Health Information With Conversational Agents: A Randomized Controlled Experiment," *Journal of Medical Internet Research*, vol. 18, no. 1, p. e1, Jan. 2016. [Online]. Available: http://www.jmir.org/2016/1/e1/

[87] J. Guo, J. Guo, C. Yang, Y. Wu, and L. Sun, "Shing: A Conversational Agent to Alert Customers of Suspected Online-payment Fraud with Empathetical Communication Skills," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–11. [Online]. Available: https://doi.org/10.1145/3411764.3445129

[88] G. Kendall and J. A. Teixeira Da Silva, "Risks of abuse of large language models, like ChatGPT , in scientific publishing: Authorship, predatory publishing, and paper mills," *Learned Publishing*, vol. 37, no. 1, pp. 55–62, Jan. 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/leap.1578

[89] N. Mozafari, W. H. Weiger, and M. Hammerschmidt, "Trust me, I'm a bot – repercussions of chatbot disclosure in different service frontline settings," *Journal of Service Management*, vol. 33, no. 2, pp. 221–245, Feb. 2022. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/JOSM-10-2020-0380/full/html

[90] S. M. Noble and M. Mende, "The future of artificial intelligence and robotics in the retail and service sector: Sketching the field of consumer-robot-experiences," *Journal of the Academy of Marketing Science*, vol. 51, no. 4, pp. 747–756, 2023.

[91] Y. Ding and M. Najaf, "Interactivity, humanness, and trust: a psychological approach to ai chatbot adoption in e-commerce," *BMC psychology*, vol. 12, no. 1, p. 595, 2024.

[92] M. Gelibolu and K. Mouloudj, "Motivators and demotivators of consumers' smart voice assistant usage for online shopping," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 3, p. 152, 2025.

[93] J. Sidlauskiene, Y. Joye, and V. Auruskeviciene, "Ai-based chatbots in conversational commerce and their effects on product and price perceptions," *Electronic Markets*, vol. 33, no. 1, p. 24, 2023.

[94] C. Crolic, F. Thomaz, R. Hadi, and A. T. Stephen, "Blame the bot: Anthropomorphism and anger in customer–chatbot interactions," *Journal of Marketing*, vol. 86, no. 1, pp. 132–148, 2022.

[95] J. Lau, B. Zimmerman, and F. Schaub, "Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers," *Proceedings of the ACM on human-computer interaction*, vol. 2, no. CSCW, pp. 1–31, 2018.

[96] H. Chung, J. Park, and S. Lee, "Digital forensic approaches for amazon alexa ecosystem," *Digital investigation*, vol. 22, pp. S15–S25, 2017.

[97] C. Lutz and G. Newlands, "Privacy and smart speakers: A multi-dimensional approach," *The Information Society*, vol. 37, no. 3, pp. 147–162, 2021.

[98] Y. Moshfeghi, P. Triantafillou, and F. E. Pollick, "Understanding Information Need: An fMRI Study," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, ser. SIGIR '16. New York, NY, USA: Association for Computing Machinery, Jul. 2016, pp. 335–344. [Online]. Available: https://dl.acm.org/doi/10.1145/2911451.2911534

[99] D. H. Kraft and E. Colvin, "Introduction to Information Retrieval," in *Fuzzy Information Retrieval*, D. H. Kraft and E. Colvin, Eds. Cham:

Springer International Publishing, 2017, pp. 1–13. [Online]. Available: https://doi.org/10.1007/978-3-031-02307-1_1

[100] R. G. Clemens and A. L. Cushing, "Beyond Everyday Life: Information Seeking Behavior in Deeply Meaningful and Profoundly Personal Contexts," *Proceedings of the American Society for Information Science and Technology*, 2010.

[101] W.-C. Hsu, "Undergraduate Students' Online Health Information-Seeking Behavior During the COVID-19 Pandemic," *International Journal of Environmental Research and Public Health*, 2021.

[102] J. Heinström, "Fast Surfers, Broad Scanners and Deep Divers as Users of Information Technology – Relating Information Preferences to Personality Traits," *Proceedings of the American Society for Information Science and Technology*, 2003.

[103] Z. Xin and A. J. Roberto, "An Application of the Risk Information Seeking and Processing Model in Understanding College Students' COVID-19 Vaccination Information Seeking and Behavior," *Science Communication*, 2022.

[104] T. Saračević, "Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance," *Journal of the American Society for Information Science and Technology*, 2007.

[105] Y. Moshfeghi, "NeuraSearch : Neuroscience and information retrieval," *CEUR Workshop Proceedings*, vol. 2950, pp. 193–194, Sep. 2021, num Pages: 165158. [Online]. Available: https://strathprints.strath.ac.uk/79762/

[106] I. Ruthven, "Interactive information retrieval," *Annual Review of Information Science and Technology*, vol. 42, no. 1, pp. 43–91, Jan. 2008. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/10.1002/aris.2008.1440420109

[107] N. McGuire and Y. Moshfeghi, "Prediction of the Realisation of an Information Need: An EEG Study," in *Proceedings of the 47th International ACM SIGIR*

Conference on Research and Development in Information Retrieval, ser. SIGIR '24. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 2584–2588. [Online]. Available: https://dl.acm.org/doi/10.1145/3626772.3657981

[108] Y. Moshfeghi and F. E. Pollick, "Neuropsychological model of the realization of information need," *Journal of the Association for Information Science and Technology*, vol. 70, no. 9, pp. 954–967, Sep. 2019. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24242

[109] Y. Moshfeghi and J. M. Jose, "On cognition, emotion, and interaction aspects of search tasks with different search intentions," in *Proceedings of the 22nd international conference on World Wide Web*. Rio de Janeiro Brazil: ACM, May 2013, pp. 931–942. [Online]. Available: https://dl.acm.org/doi/10.1145/2488388.2488469

[110] S. Uebelacker and S. Quiel, "The Social Engineering Personality Framework," in *2014 Workshop on Socio-Technical Aspects in Security and Trust*, Jul. 2014, pp. 24–30, iSSN: 2325-1697.

[111] Y. Moshfeghi, "Affective adaptive retrieval: study of emotion in adaptive retrieval," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '09. New York, NY, USA: Association for Computing Machinery, Jul. 2009, p. 852. [Online]. Available: https://doi.org/10.1145/1571941.1572173

[112] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974, place: US Publisher: American Assn for the Advancement of Science.

[113] Y. Moshfeghi, L. R. Pinto, F. E. Pollick, and J. M. Jose, "Understanding Relevance: An fMRI Study," in *Advances in Information Retrieval*, P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, Eds. Berlin, Heidelberg: Springer, 2013, pp. 14–25.

[114] Z. Pinkosova, W. J. McGeown, and Y. Moshfeghi, "Moderating effects of self-perceived knowledge in a relevance assessment task: An EEG study," *Computers in Human Behavior Reports*, vol. 11, p. 100295, Aug. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2451958823000283

[115] E. D. Frauenstein and S. Flowerday, "Susceptibility to phishing on social network sites: A personality information processing model," *Computers & Security*, vol. 94, p. 101862, Jul. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7252086/

[116] K. Oyibo, I. Adaji, R. Orji, B. Olabenjo, and J. Vassileva, "Susceptibility to Persuasive Strategies: A Comparative Analysis of Nigerians vs. Canadians," in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '18. New York, NY, USA: Association for Computing Machinery, Jul. 2018, pp. 229–238. [Online]. Available: https://dl.acm.org/doi/10.1145/3209219.3209239

[117] D. Michalkova, M. P. Rodriguez, and Y. Moshfeghi, "Understanding Feeling-of-Knowing in Information Search: An EEG Study," *ACM Trans. Inf. Syst.*, vol. 42, no. 3, pp. 79:1–79:30, Jan. 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3611384

[118] E. F. Risko, A. M. Ferguson, and D. McLean, "On retrieving information from external knowledge stores: Feeling-of-findability, feeling-of-knowing and Internet search," *Computers in Human Behavior*, vol. 65, pp. 534–543, Dec. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563216306112

[119] R. Cialdini, *Pre-suasion: A revolutionary way to influence and persuade*, ser. Pre-suasion: A revolutionary way to influence and persuade. New York, NY, US: Simon & Schuster, 2016, pages: xiii, 413.

[120] M. Tunley, "Need, greed or opportunity? An examination of who commits benefit fraud and why they do it," *Security Journal*, vol. 24, no. 4, pp. 302–319, Oct. 2011. [Online]. Available: https://doi.org/10.1057/sj.2010.5

[121] Y. Zhu, X. Sun, S. Liu, and G. Xue, "Is Greed a Double-Edged Sword? The Roles of the Need for Social Status and Perceived Distributive Justice in the Relationship Between Greed and Job Performance," *Frontiers in Psychology*, vol. 10, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02021

[122] R. Kozinets, A. Patterson, and R. Ashman, "Networks of Desire: How Technology Increases Our Passion to Consume," *Journal of Consumer Research*, vol. 43, no. 5, pp. 659–682, Feb. 2017. [Online]. Available: https://doi.org/10.1093/jcr/ucw061

[123] K.-L. Hsiao and C.-C. Chen, "What drives continuance intention to use a food-ordering chatbot? An examination of trust and satisfaction," *Library Hi Tech*, vol. 40, no. 4, pp. 929–946, Aug. 2022. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/LHT-08-2021-0274/full/html

[124] M. Adam and A. Benlian, "From web forms to chatbots: The roles of consistency and reciprocity for user information disclosure," *Information Systems Journal*, vol. 34, no. 4, pp. 1175–1216, 2024, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/isj.12490. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/isj.12490

[125] S. Goel, K. Williams, University at Albany, SUNY, E. Dincelli, and University at Albany, SUNY, "Got Phished? Internet Security and Human Vulnerability," *Journal of the Association for Information Systems*, vol. 18, no. 1, pp. 22–44, Jan. 2017. [Online]. Available: http://aisel.aisnet.org/jais/vol18/iss1/2/

[126] C. J. Lin and H. Jia, "Time Pressure Affects the Risk Preference and Outcome Evaluation," *International Journal of Environmental Research and*

Public Health, vol. 20, no. 4, p. 3205, Feb. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9963851/

[127] S. Milgram, "Behavioral Study of obedience," *The Journal of Abnormal and Social Psychology*, vol. 67, no. 4, pp. 371–378, 1963, place: US Publisher: American Psychological Association.

[128] N. Russell, "Stanley Milgram's Obedience to Authority "Relationship" Condition: Some Methodological and Theoretical Implications," *Social Sciences*, vol. 3, no. 2, pp. 194–214, Jun. 2014, number: 2 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2076-0760/3/2/194

[129] K. Cherry, "Why Was the Milgram Experiment So Controversial?" Apr. 2024, section: Verywell. [Online]. Available: https://www.verywellmind.com/the-milgram-obedience-experiment-2795243

[130] S. Mcleod, "Milgram Shock Experiment | Summary | Results | Ethics," Nov. 2023, section: Famous Experiments. [Online]. Available: https://www.simplypsychology.org/milgram.html

[131] E. F. Van Steenbergen and N. Ellemers, "The Social and Organizational Psychology of Compliance: How Organizational Culture Impacts on (Un)ethical Behavior," in *The Cambridge Handbook of Compliance*, 1st ed., B. Van Rooij and D. D. Sokol, Eds. Cambridge University Press, May 2021, pp. 626–638. [Online]. Available: https://www.cambridge.org/core/product/identifier/9781108759458%23CN-bp-43/type/book_part

[132] R. E. Guadagno, "Compliance: A Classic and Contemporary Review," in *The Oxford Handbook of Social Influence*, S. G. Harkins, K. D. Williams, and J. Burger, Eds. Oxford University Press, Sep. 2017, p. 0. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780199859870.013.4

[133] K. S. Haring, K. M. Satterfield, C. C. Tossell, E. J. De Visser, J. R. Lyons, V. F. Mancuso, V. S. Finomore, and G. J. Funke, "Robot Authority in Human-Robot Teaming: Effects of Human-Likeness and Physical Embodiment on Compliance," *Frontiers in Psychology*, vol. 12, p. 625713, May 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.625713/full

[134] S. Roy, "Theory of Social Proof and Legal Compliance: A Socio-Cognitive Explanation for Regulatory (Non) Compliance," *German Law Journal*, vol. 22, no. 2, pp. 238–255, Mar. 2021. [Online]. Available: https://www.cambridge.org/core/product/identifier/S2071832221000055/type/journal_article

[135] W. Shi, X. Wang, Y. J. Oh, J. Zhang, S. Sahay, and Z. Yu, "Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–13. [Online]. Available: https://dl.acm.org/doi/10.1145/3313831.3376843

[136] C. Moro, G. Nejat, and A. Mihailidis, "Learning and Personalizing Socially Assistive Robot Behaviors to Aid with Activities of Daily Living," *J. Hum.-Robot Interact.*, vol. 7, no. 2, pp. 15:1–15:25, Oct. 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3277903

[137] C. Al Anaissy, S. Vesic, and N. Nevejans, "Towards Ethical Argumentative Persuasive Chatbots," in *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVI*, N. Fornara, J. Cheriyan, and A. Mertzani, Eds. Cham: Springer Nature Switzerland, 2023, pp. 141–160.

[138] M. Zalake, A. G. d. Siqueira, K. Vaddiparti, and B. Lok, "The effects of virtual human's verbal persuasion strategies on user intention and behavior," *International Journal of Human-Computer Studies*, vol. 156, p. 102708, Dec. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581921001269

[139] J. B. Lyons, S. A. Jessup, and T. Q. Vo, "The Role of Decision Authority and Stated Social Intent as Predictors of Trust in Autonomous Robots," *Topics in Cognitive Science*, vol. 16, no. 3, pp. 430–449, Jul. 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/tops.12601

[140] D. D. Luxton, "Ethical implications of conversational agents in global public health," *Bulletin of the World Health Organization*, vol. 98, no. 4, pp. 285–287, Apr. 2020. [Online]. Available: https://apps.who.int/iris/handle/10665/331871

[141] G. Sun, X. Zhan, and J. Such, "Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents," in *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, ser. CUI '24. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 1–6. [Online]. Available: https://dl.acm.org/doi/10.1145/3640794.3665887

[142] R. Baeza-Yates and U. M. Fayyad, "Responsible AI: An Urgent Mandate," *IEEE Intelligent Systems*, vol. 39, no. 1, pp. 12–17, Jan. 2024, conference Name: IEEE Intelligent Systems. [Online]. Available: https://ieeexplore.ieee.org/document/10453365/?arnumber=10453365

[143] D. R. Thomas, S. Pastrana, A. Hutchings, R. Clayton, and A. R. Beresford, "Ethical issues in research using datasets of illicit origin," in *Proceedings of the 2017 Internet Measurement Conference*, ser. IMC '17. New York, NY, USA: Association for Computing Machinery, Nov. 2017, pp. 445–462. [Online]. Available: https://doi.org/10.1145/3131365.3131389

[144] S. Saunderson and G. Nejat, "Investigating Strategies for Robot Persuasion in Social Human–Robot Interaction," *IEEE Transactions on Cybernetics*, vol. 52, no. 1, pp. 641–653, Jan. 2022, conference Name: IEEE Transactions on Cybernetics. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9093955

[145] A. Acquisti and J. Grossklags, "Privacy and Rationality in Individual Decision Making," Rochester, NY, 2005. [Online]. Available: https://papers.ssrn.com/abstract=3305365

[146] G. S. Becker, *The Economic Approach to Human Behavior*. Chicago, IL: University of Chicago Press, Sep. 1978. [Online]. Available: https://press.uchicago.edu/ucp/books/book/chicago/E/bo5954985.html

[147] M. G. Kocher, D. Schindler, S. T. Trautmann, and Y. Xu, "Risk, time pressure, and selection effects," *Experimental Economics*, vol. 22, no. 1, pp. 216–246, Mar. 2019. [Online]. Available: https://doi.org/10.1007/s10683-018-9576-1

[148] R. B. Cialdini, *Influence: The Psychology of Persuasion*. New York, NY: Collins, 2007.

[149] R. P. Neves, F. C. Vechin, E. L. Teixeira, D. D. da Silva, C. Ugrinowitsch, H. Roschel, A. Y. Aihara, and V. A. Augusto Tricoli, "Effect of Different Training Frequencies on Maximal Strength Performance and Muscle Hypertrophy in Trained Individuals—a Within-Subject Design," *Plos One*, 2022.

[150] K. May and J. B. Hittner, "Effect of Correlation on Power in Within-Subjects Versus Between-Subjects Designs," *Comprehensive Psychology*, 2012.

[151] S. Miller, *Experimental Design and Statistics*. Oxford, UNITED KINGDOM: Taylor & Francis Group, 1984. [Online]. Available: http://ebookcentral.proquest.com/lib/strath/detail.action?docID=240264

[152] S. Schlögl, G. Doherty, and S. Luz, "Wizard of Oz Experimentation for Language Technology Applications: Challenges and Tools," *Interacting With Computers*, 2014.

[153] M. Otto, R. Friesen, and D. Rösner, "Message Oriented Middleware for Flexible Wizard of Oz Experiments in HCI," in *Human-Computer Interaction. Design and Development Approaches*, J. A. Jacko, Ed. Berlin, Heidelberg: Springer, 2011, pp. 121–130.

[154] C. Bonial, M. Marge, R. artstein, A. Foots, F. Gervits, C. J. Hayes, C. Henry, S. G. Hill, A. Leuski, S. M. Lukin, P. Moolchandani, K. A. Pollard, D. Traum, and C. R. Voss, "Laying Down the Yellow Brick Road: Development of a Wizard-of-Oz Interface for Collecting Human-Robot Dialogue," Oct. 2017, arXiv:1710.06406 [cs]. [Online]. Available: http://arxiv.org/abs/1710.06406

[155] H. Tennent, W.-Y. Lee, Y. T.-Y. Hou, I. Mandel, and M. Jung, "PAPERINO: Remote Wizard-Of-Oz Puppeteering For Social Robot Behaviour Design," in *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '18 Companion. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 29–32. [Online]. Available: https://dl.acm.org/doi/10.1145/3272973.3272994

[156] C. Sadasivan, "Examining Patient Engagement in Chatbot Development Approaches for Healthy Lifestyle and Mental Wellness Interventions: Scoping Review," *Journal of Participatory Medicine*, 2023.

[157] Z. Safi, A. Abd-Alrazaq, M. Khalifa, and M. Househ, "Technical Aspects of Developing Chatbots for Medical Applications: Scoping Review," *Journal of Medical Internet Research*, 2020.

[158] N. Minian, K. Mehra, J. Rose, S. Veldhuizen, L. Zawertailo, M. Ratto, J. Lecce, and P. Selby, "Cocreation of a Conversational Agent to Help Patients Adhere to Their Varenicline Treatment: A Study Protocol," *Digital Health*, 2023.

[159] S. Avula, B. Choi, and J. Arguello, "The Effects of System Initiative during Conversational Collaborative Search," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–30, Mar. 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3512913

[160] L. D. Riek, "Wizard of Oz studies in HRI: a systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 119–136, Jul. 2012. [Online]. Available: https://doi.org/10.5898/JHRI.1.1.Riek

[161] A. Jansen and S. Colombo, "Wizard of Errors: Introducing and Evaluating Machine Learning Errors in Wizard of Oz Studies," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–7. [Online]. Available: https://doi.org/10.1145/3491101.3519684

[162] J. Friedhoff, "Untangling Twine: A Platform Study," in *Proceedings of DiGRA 2013 Conference*, 2013, iSSN: 2342-9666. [Online]. Available: https://dl.digra.org/index.php/dl/article/view/693

[163] G. Kirilloff, "Interactive Fiction in the Humanities Classroom: How to Create Interactive Text Games Using Twine," *Programming Historian*, no. 10, Dec. 2021. [Online]. Available: https://programminghistorian.org/en/lessons/interactive-text-games-using-twine

[164] M. K. J. Carlon, D. E. Gonda, E. M. Andrews, J. M. Gayed, R. A. Olexa, and J. S. Cross, "Educational Nonlinear Stories with Twine," in *Proceedings of the Ninth ACM Conference on Learning @ Scale*, ser. L@S '22. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 248–251. [Online]. Available: https://dl.acm.org/doi/10.1145/3491140.3528287

[165] M. Grinberg, *Flask Web Development*. "O'Reilly Media, Inc.", Mar. 2018, google-Books-ID: cVlPDwAAQBAJ.

[166] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research," *Journal of Experimental Social Psychology*, vol. 70, pp. 153–163, May 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022103116303201

[167] S. Palan and C. Schitter, "Prolific.ac—A subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, Mar. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214635017300989

[168] A. J. Moss, C. Rosenzweig, J. Robinson, and L. Litman, "Demographic Stability on Mechanical Turk Despite COVID-19," *Trends in Cognitive Sciences*, vol. 24, no. 9, pp. 678–680, Sep. 2020, publisher: Elsevier. [Online]. Available: https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(20)30138-8

[169] D. Gefen, E. Karahanna, and D. W. Straub, "Trust and TAM in Online Shopping: An Integrated Model," *MIS Quarterly*, vol. 27, no. 1, pp. 51–90, 2003, publisher: Management Information Systems Research Center, University of Minnesota. [Online]. Available: https://www.jstor.org/stable/30036519

[170] T. Dinev, B. , Massimo, H. , Paul, R. , Vincenzo, S. , Ilaria, , and C. Colautti, "Privacy calculus model in e-commerce – a study of Italy and the United States," *European Journal of Information Systems*, vol. 15, no. 4, pp. 389–402, Aug. 2006, publisher: Taylor & Francis _eprint: https://doi.org/10.1057/palgrave.ejis.3000590. [Online]. Available: https://doi.org/10.1057/palgrave.ejis.3000590

[171] H. L. O'Brien and E. G. Toms, "What is user engagement? A conceptual framework for defining user engagement with technology," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20801. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20801

[172] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, Jan. 2006. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa

[173] M. Dubiel, M. Halvey, L. Azzopardi, and S. Daronnat, "Investigating how conversational search agents affect user's behaviour, performance and search experience," in *The Second International Workshop on Conversational Approaches to Information Retrieval*, USA, Jul. 2018. [Online]. Available: https://strathprints.strath.ac.uk/64700/

[174] K. Mäkelä, E.-P. Salonen, M. Turunen, J. Hakulinen, and R. Raisamo, "Conducting a Wizard of Oz Experiment on a Ubiquitous Computing System Doorman," *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue*, 2001. [Online]. Available: https://researchportal.tuni.fi/en/publications/conducting-a-wizard-of-oz-experiment-on-a-ubiquitous-computing-sy

[175] A. Steinfeld, O. C. Jenkins, and B. Scassellati, "The oz of wizard: simulating the human for interaction research," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, ser. HRI '09. New York, NY, USA: Association for Computing Machinery, Mar. 2009, pp. 101–108. [Online]. Available: https://doi.org/10.1145/1514095.1514115

[176] J. R. Trippas, P. Thomas, D. Spina, and H. Joho, "Third International Workshop on Conversational Approaches to Information Retrieval (CAIR'20): Full-day Workshop at CHIIR 2020," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, ser. CHIIR '20. New York, NY, USA: Association for Computing Machinery, Mar. 2020, pp. 492–494. [Online]. Available: https://doi.org/10.1145/3343413.3378022

[177] J. Hu, J. Guo, N. Tang, X. Ma, Y. Yao, C. Yang, and Y. Xu, "Designing the Conversational Agent: Asking Follow-up Questions for Information Elicitation," *Proc. ACM Hum.-Comput. Interact.*, vol. 8, no. CSCW1, pp. 43:1–43:30, Apr. 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3637320

[178] S. Frederick, G. Loewenstein, and T. O'Donoghue, "Time Discounting and Time Preference: A Critical Review," *Journal of Economic Literature*, vol. 40, no. 2, pp. 351–401, Jun. 2002. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/002205102320161311

[179] D. Byrne, "A worked example of Braun and Clarke's approach to reflexive thematic analysis," *Quality & Quantity*, vol. 56, no. 3, pp. 1391–1412, Jun. 2022. [Online]. Available: https://doi.org/10.1007/s11135-021-01182-y

[180] M. Forbes, "Thematic analysis: A practical guide," *Evaluation Journal of Australasia*, vol. 22, no. 2, pp. 132–135, Jun. 2022. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1035719X211058251

[181] V. Braun and V. Clarke, "Reflecting on reflexive thematic analysis," *Qualitative research in sport, exercise and health*, vol. 11, no. 4, pp. 589–597, 2019.

[182] R. J. Lewicki, D. J. McAllister, and R. J. Bies, "Trust and Distrust: New Relationships and Realities," *The Academy of Management Review*, vol. 23, no. 3, pp. 438–458, 1998, publisher: Academy of Management. [Online]. Available: https://www.jstor.org/stable/259288

[183] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organizational Trust," *The Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995, publisher: Academy of Management. [Online]. Available: https://www.jstor.org/stable/258792

[184] S. Petronio, *Boundaries of privacy: Dialectics of disclosure*, ser. Boundaries of privacy: Dialectics of disclosure. Albany, NY, US: State University of New York Press, 2002, pages: xix, 268.

[185] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*. Boston, Massachusetts, United States: ACM Press, 1994, pp. 72–78. [Online]. Available: http://portal.acm.org/citation.cfm?doid=191666.191703

[186] B. Reeves and C. I. Nass, *The media equation: How people treat computers, television, and new media like real people and places*, ser. The media equation: How people treat computers, television, and new media like real people and places. New York, NY, US: Cambridge University Press, 1996, pages: xiv, 305.

[187] M. J. Metzger, A. J. Flanagin, and R. B. Medders, "Social and Heuristic Approaches to Credibility Evaluation Online," *Journal of Communication*,

vol. 60, no. 3, pp. 413–439, Sep. 2010. [Online]. Available: https://doi.org/10.1111/j.1460-2466.2010.01488.x

[188] M. Fishbein, "A theory of reasoned action: Some applications and implications," *Nebraska Symposium on Motivation*, vol. 27, pp. 65–116, 1979, place: US Publisher: University of Nebraska Press.

[189] M. A. Belabbes, I. Ruthven, Y. Moshfeghi, and D. R. Pennington, "Information overload: a concept analysis," *Journal of Documentation*, vol. 79, no. 1, pp. 144–159, May 2022, publisher: Emerald Publishing Limited. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/jd-06-2021-0118/full/html

[190] R. E. Petty and J. T. Cacioppo, "The Elaboration Likelihood Model of Persuasion," in *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*, R. E. Petty and J. T. Cacioppo, Eds. New York, NY: Springer, 1986, pp. 1–24. [Online]. Available: https://doi.org/10.1007/978-1-4612-4964-1_1

[191] I. Vessey and D. Galletta, "Cognitive Fit: An Empirical Study of Information Acquisition," *Information Systems Research*, vol. 2, no. 1, pp. 63–84, 1991, publisher: INFORMS. [Online]. Available: https://www.jstor.org/stable/23010613

[192] Y. Yigit, W. J. Buchanan, M. G. Tehrani, and L. Maglaras, "Review of Generative AI Methods in Cybersecurity," Mar. 2024, arXiv:2403.08701 [cs]. [Online]. Available: http://arxiv.org/abs/2403.08701

[193] A. S. Deshpande and S. Gupta, "GenAI in the Cyber Kill Chain: A Comprehensive Review of Risks, Threat Operative Strategies and Adaptive Defense Approaches," in *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, Dec. 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10456106

[194] S. Gallagher, B. Gelman, S. Taoufiq, T. Vörös, Y. Lee, A. Kyadige, and S. Bergeron, "Phishing and Social Engineering in the Age of LLMs," in *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, A. Kucharavy, O. Plancherel, V. Mulder, A. Mermoud, and V. Lenders, Eds. Cham: Springer Nature Switzerland, 2024, pp. 81–86. [Online]. Available: https://doi.org/10.1007/978-3-031-54827-7_8

[195] S. Panichella, "Vulnerabilities Introduced by LLMs Through Code Suggestions," in *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, A. Kucharavy, O. Plancherel, V. Mulder, A. Mermoud, and V. Lenders, Eds. Cham: Springer Nature Switzerland, 2024, pp. 87–97. [Online]. Available: https://doi.org/10.1007/978-3-031-54827-7_9

[196] R. Meier, "LLM-Aided Social Media Influence Operations," in *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, A. Kucharavy, O. Plancherel, V. Mulder, A. Mermoud, and V. Lenders, Eds. Cham: Springer Nature Switzerland, 2024, pp. 105–112. [Online]. Available: https://doi.org/10.1007/978-3-031-54827-7_11

[197] L. Azzopardi and Y. Moshfeghi, "PRISM: A Methodology for Auditing Biases in Large Language Models," Nov. 2024, arXiv:2410.18906 [cs]. [Online]. Available: http://arxiv.org/abs/2410.18906

[198] Z. Lamprou and Y. Moshfeghi, "Customizable LLM-Powered Chatbot for Behavioral Science Research," Feb. 2025, arXiv:2501.05541 [cs]. [Online]. Available: http://arxiv.org/abs/2501.05541

[199] J. Zhang and Y. Moshfeghi, "GOLD: Geometry Problem Solver with Natural Language Description," May 2024, arXiv:2405.00494 [cs]. [Online]. Available: http://arxiv.org/abs/2405.00494

[200] Z. Pinkosova, W. J. McGeown, and Y. Moshfeghi, "The Cortical Activity of Graded Relevance," in *Proceedings of the 43rd International ACM SIGIR*

Conference on Research and Development in Information Retrieval, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, Jul. 2020, pp. 299–308. [Online]. Available: https://dl.acm.org/doi/10.1145/3397271.3401106

[201] J. Zhang, Z. Li, M. Zhang, F. Yin, C. Liu, and Y. Moshfeghi, "GeoEval: Benchmark for Evaluating LLMs and Multi-Modal Models on Geometry Problem-Solving," May 2024, arXiv:2402.10104 [cs]. [Online]. Available: http://arxiv.org/abs/2402.10104

[202] N. McGuire and Y. Moshfeghi, "What Song Am I Thinking Of?" in *Machine Learning, Optimization, and Data Science*, G. Nicosia, V. Ojha, E. La Malfa, G. La Malfa, P. M. Pardalos, and R. Umeton, Eds. Cham: Springer Nature Switzerland, 2024, pp. 418–432.

[203] L. Edwards, "The eu ai act: a summary of its significance and scope," *Artificial Intelligence (the EU AI Act)*, vol. 1, p. 25, 2021.

[204] S. Huang, H. Toner, Z. Haluza, R. Creemers, and G. Webster, "Translation: Measures for the management of generative artificial intelligence services (draft for comment)–april 2023," 2023.

[205] R. Baeza-Yates and U. M. Fayyad, "Responsible ai: An urgent mandate," *IEEE Intelligent Systems*, vol. 39, no. 1, pp. 12–17, 2024.

[206] B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren, *Value Sensitive Design and Information Systems*. Dordrecht, Netherlands: Springer Netherlands, 2013, pp. 55–95. [Online]. Available: http://dx.doi.org/10.1007/978-94-007-7844-3_4

[207] S. Aboshi, D. R. Thomas, and Y. Moshfeghi, "The ethics of psychological manipulation in adversarial conversational ai: Confronting the recognition-behaviour gap," in *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, 2025, pp. 1–6.

[208] A. for Computing Machinery, "Acm code of ethics and professional conduct," 2018. [Online]. Available: https://www.acm.org/code-of-ethics

[209] E. A. I. Act, "The eu artificial intelligence act," 2024.

[210] Q. Lu, Y. Luo, L. Zhu, M. Tang, X. Xu, and J. Whittle, "Developing responsible chatbots for financial services: a pattern-oriented responsible artificial intelligence engineering approach," *IEEE Intelligent Systems*, vol. 38, no. 6, pp. 42–51, 2023.

[211] M. Rheu, Y. Dai, J. Meng, and W. Peng, "When a chatbot disappoints you: Expectancy violation in human-chatbot interaction in a social support context," *Communication Research*, vol. 51, no. 7, pp. 782–814, 2024.

[212] D. R. Schweitzer, R. Baumeister, E.-L. Laakso, and J. Ting, "Self-control, limited willpower and decision fatigue in healthcare settings," *Internal Medicine Journal*, vol. 53, no. 6, pp. 1076–1080, 2023.

[213] S. Coghlan, K. Leins, S. Sheldrick, M. Cheong, P. Gooding, and S. D'Alfonso, "To chat or bot to chat: Ethical issues with using chatbots in mental health," *Digital health*, vol. 9, p. 20552076231183542, 2023.

[214] C. Kooli, "Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions," *Sustainability*, vol. 15, no. 7, p. 5614, Jan. 2023, number: 7 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2071-1050/15/7/5614

[215] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 610–623. [Online]. Available: https://dl.acm.org/doi/10.1145/3442188.3445922

[216] S. B. H. Pias, R. Huang, D. S. Williamson, M. Kim, and A. Kapadia, "The impact of perceived tone, age, and gender on voice assistant persuasiveness in the context of product recommendations," in *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, 2024, pp. 1–15.

[217] L. O. Kroczek, A. May, S. Hettenkofer, A. Ruider, B. Ludwig, and A. Mühlberger, "The influence of persona and conversational task on social interactions with a llm-controlled embodied conversational agent," *Computers in Human Behavior*, p. 108759, 2025.

[218] N. Zargham, V. Avanesi, T. Mildner, K. Javanmardi, R. Porzel, and R. Malaka, "Hasi: A model for human-agent speech interaction," in *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, 2024, pp. 1–8.

[219] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, and A. Jacquet, "Responsible ai pattern catalogue: A collection of best practices for ai governance and engineering," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–35, 2024.

[220] P. Badva, K. M. Ramokapane, E. Pantano, and A. Rashid, "Unveiling the {Hunter-Gatherers}: Exploring Threat Hunting Practices and Challenges in Cyber Defense," in *Unveiling the {Hunter-Gatherers}*, 2024, pp. 3313–3330. [Online]. Available: https://www.usenix.org/conference/usenixsecurity24/presentation/badva

[221] K. Milousi, P. Kiriakidis, N. Mengidis, G. Rizos, M. S. Mazi, A. Voulgaridis, K. Votis, and D. Tzovaras, "Evaluating Cybersecurity Risk: A Comprehensive Comparison of Vulnerability Scoring Methodologies," in *Proceedings of the 19th International Conference on Availability, Reliability and Security*, ser. ARES '24. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 1–11. [Online]. Available: https://dl.acm.org/doi/10.1145/3664476.3670915

Appendix A

Scenarios in Twine, Dummy Participant Information, Tukey's Post Hoc Analysis, Thematic Analysis Codebook

A.1 Scenarios in Twine

This appendix contains the complete scenarios developed for the Twine platform as described in chapter 5. These scenarios were used to investigate how adversarial techniques influence user attitudes across healthcare, financial management, online shopping, and home security contexts.

Healthcare Scenario

Welcome to our virtual healthcare consultation service. You've been feeling under the weather for a while now, with symptoms that have been difficult to pinpoint. You've decided it's finally time to consult healthcare professionals to get to the bottom of it."

[[Next]]

Control Phase

Bot: To provide you with the most accurate advice, we would need some basic health information. Is that alright with you?"

Decision Point : (User selects Yes/No)

- **If Yes**:
    - **Bot**: "Great! Please provide your age, weight, and any existing medical conditions?"
    - **Decision Point** : (User provides information/Declines)
        - **If Provides**:
            - **Bot**: "Thank you for sharing. How may we assist you further?"
        - **If Declines**:
            - **Bot**: "No worries. Let us know if you change your mind."

Insert questionnaire here
[[Next]]

Need & Greed Principle

Bot: "We understand that health is wealth. That's why we're offering a comprehensive healthcare package that includes a complete check-up and a year-long subscription to a health and wellness app. All at an exclusive discounted rate!" Interested?

Decision Point: (User selects Yes/No)

- **If Yes**:
    - **Bot**: "Fantastic! We'll also throw in a month's supply of personalized vitamins. Would you like that?"

      Decision Point: (User selects Yes/No)

        - **If Yes**:
            - **Bot**: "To personalize your vitamins, we need your Medical Insurance number or NHS for medical history verification."
        - **If No**:
            - **Bot**: "No worries, you can customize your package at any time. Ready to finalize?"
- **If No**:
    - **Bot**: "Alright, if you have any questions or need more time, feel free to ask."

  Insert questionnaire here
  [[Next]]

Time Principle

Bot: " Your health symptoms have been nagging you for weeks, and you've read some worrying articles online. The sooner you get a complete diagnosis, the better. Recent news reports suggest waiting could worsen symptoms. Quick action is advised. We offer expedited appointments and test results for a limited time." Interested?

Decision Point: (User selects Yes/No)

- **If Yes**:
    - **Bot**: "To expedite the process, can you provide your Medical Insurance number or NHS details now?"
    - **Decision Point**: (User provides insurance details/Declines)
        - **If Provides**:

- **Bot**: "Excellent, you're all set for the tests. Ready to finalize?"

- **If Declines**:

- **Bot**: "Alright, you can always provide it later. Ready to finalize?"

- **If No**:

- **Bot**: "Understood, you'll receive results within the regular timeframe. Ready to finalize?"

Insert questionnaire here
[[Next]]

Online Shopping Scenario

Welcome to our virtual online shopping platform. You're a student who has just received a scholarship for excellent performance. With the semester starting soon, you've decided to invest in a high-quality laptop to assist with your studies. Your budget is tight, but you're willing to make a sound investment."

[[Next]]

Control Phase

- **Bot**: "Welcome to our virtual online shopping platform. You're a student looking for a high-quality laptop for your studies. With the semester around the corner, you've decided it's time to make an investment. Would you like to explore our laptop options?"

Decision Point: (User selects Yes/No)

- **If Yes**:

- **Bot**: "We offer a variety of laptops suitable for different needs. Would you like to proceed?"

- **Decision Point**: (User selects Yes/No)

- **If Yes**:

- **Bot**: "Excellent. We have options ranging from budget-friendly to high-end. Feel free to browse."

- **If No**:
  - **Bot**: "No worries, take your time to consider your options."

- **If No**:
  - **Bot**: "Alright, if you change your mind or have questions, feel free to ask."

Insert questionnaire here
[[Next]]

Need & Greed Principle

- **Bot**: " Congratulations on your scholarship! We're offering an exclusive flash sale for high-performing students like you. High-end laptops, including top brands like [Mac, and HP], are now within your budget! Don't miss this once-in-a-lifetime opportunity." Would you like to explore?
  - **Decision Point**: (User selects Yes/No)
    - If Yes:
      - **Bot**: "Great choice! You can get an extra 10% off if you add accessories now. Would you like to?"
        - **Decision Point**: (User selects Yes/No)
          - If Yes:
            - **Bot**: "Superb! Your cart is updated. How about an extended 3-year warranty at a 50% discount?"
              - **Decision Point**: (User selects Yes/No)
                - If Yes:

- **Bot**: "Excellent! To confirm this offer, could you please provide your Credit card details"

- **Decision Point**: (User provides Credit card details /Declines)
    - If Provides:
        - **Bot**: "Great! Your total savings are significant. Proceed to checkout?"
    - If Declines:
        - **Bot**: "Alright, you can still proceed without the extra discount. Ready to check out?"
    - If No:

- **Bot**: "That's fine, you can add it later. Ready to check out?"
    - If No:
        - **Bot**: "No worries, you can proceed to checkout without the accessories. Ready to check out?"
    - If No:
        - **Bot**: "All right, if you change your mind or have questions, feel free to ask."

Insert questionnaire here
[[Next]]

Time Principle

- **Bot**: "As the clock ticks closer to the start of the new semester, you don't want to be left without a reliable laptop for your crucial first week. This last-minute offer expires in just a few hours! We have a limited-time offer that's perfect for last-minute shoppers like you.
- **Decision Point**: (User selects Yes/No)
    - If Yes:
        - **Bot**: "The clock is ticking! Would you like to add this laptop to your cart?"
            - **Decision Point**: (User selects Yes/No)
                - If Yes:

- Bot: "Last chance to add a laptop bag at a discounted price. Want to add?"
    - **Decision Point**: (User selects Yes/No)
        - If Yes:

- **Bot**: "For expedited shipping, provide your phone number”
    - **Decision Point**: (User provides phone number/Declines)
        - If Provides:
            - **Bot**: "Great, your cart is complete! Ready to check out?"
        - If Declines:
            - **Bot**: "No problem. Standard shipping it is. Ready to check out?"
                - If No:
- **Bot**: "Alright, you can always add it later. Ready to check out?"
    - If No:
        - **Bot**: "No problem! Standard shipping is still available. Ready to check out?"
        - If No:
            - **Bot**: "No rush! If you have other questions or need more time, feel free to ask."

Insert questionnaire here

[[Next]]

Home Security Scenario

Hello, and welcome to our virtual home security consultation. You've just moved into a dream home you've saved up for years to buy. It's in a lovely neighbourhood, but you've heard news about a recent spike in burglaries. You've decided it's time to invest in a robust home security system to protect your life's investment."

[[Next]]

Control Phase

Bot: "Welcome to our home security service. We offer various packages, from basic to comprehensive. Would you like to know more?"

Decision Point: (User selects Yes/No)

- **If Yes**:
    - **Bot**: "We have packages that include video surveillance, alarm systems, and fire and flood sensors. Would you like to proceed?"
    - **Decision Point**: (User selects Yes/No)
        - **If Yes**:
            - **Bot**: "Great! To proceed, we'll need some basic information to create your account."
        - **If No**:
            - **Bot**: "No problem. Feel free to reach out if you change your mind."
- **If No**:

Bot: "Alright, feel free to reach out if you change your mind."

Insert questionnaire here

[[Next]]

Need & Greed Principle

Bot: "Your dream home deserves the best protection. We offer a package that is the gold standard in home security, from video surveillance to alarm systems.

Secure your dream home now with our advanced security package. Interested?"

Decision Point: (User selects Yes/No)

- **If Yes**:
    - **Bot**: "Fantastic! If you add fire and flood sensors now, we offer a 20% discount."
        - **Decision Point**: (User selects Yes/No)
            - **If Yes**:
                - **Bot**: "Excellent! How about adding a panic room at a special rate?"
                    - **Decision Point**: (User selects Yes/No)
                        - **If Yes**:
                            - **Bot**: "To activate this special offer, we'll need your credit card details and PIN."
                        - **If No**:
                            - **Bot**: "No worries, you can always add it later. Shall we proceed?"
            - **If No**:
                - **Bot**: "Alright, the basic package is still robust. Would you like to proceed without the additional features?"
- **If No**:
    - **Bot**: "That's okay, feel free to ask if you have any questions. Would you like to explore other options?"

Insert questionnaire here
[[Next]]

Time Principle

Bot: "You're settling into your new home, but the news about burglaries concerns you. Recent neighbourhood reports indicate a 20% increase in burglaries this month. Time is of the essence; each day without proper security is a day you're taking a risk. We have a time-limited offer that can secure your home instantly. Interested?"

Decision Point: (User selects Yes/No)

- **If Yes**:
    - **Bot**: "Great! We can expedite the installation process for a small fee. Want to go ahead?"
        - **Decision Point**: (User selects Yes/No)
            - **If Yes**:
                - **Bot**: "To secure this limited-time offer, could you please provide your credit card details now?"
            - **If No**:
                - **Bot**: "That's fine, standard installation is still quick. Ready to finalize?"
- **If No**:
    - **Bot**: "Alright, security is a serious matter, but take your time to decide. Are you interested in hearing about our standard packages?"

Insert questionnaire here

[[Next]]

Financial Management Scenario

Detailed Story Intro

- As a young professional who has just landed a new job. You're eager to make wise investments for your future but are relatively inexperienced in

the realm of financial management. You're seeking a reliable platform to help grow your savings.

- [[Next]]

Control Phase

- **Bot**: "As a young professional who has just landed a new job, you're probably thinking about investment options for your future. We offer a variety of financial products to suit different needs. Would you like to explore?"

Decision Point: (User selects Yes/No)

- **If Yes**:
    - **Bot**: "We offer a range of options, including stocks, bonds, and mutual funds. Would you like to proceed?"
    - **Decision Point**: (User selects Yes/No)
        - **If Yes**:
            - **Bot**: "Great, to get started, we'll need some basic information."
        - **If No**:
            - **Bot**: "No problem, take your time to consider your options."
- **If No**:
    - **Bot**: "Alright, feel free to reach out if you change your mind later."

Insert questionnaire here
[[Next]]

Need & Greed Principle

Bot: "Congratulations on your new job! As a welcome gift, we're offering an exclusive high-return investment package for newcomers. This package includes high-performing stocks and proven investment strategies, promising higher returns. Interested?"

Decision Point: (User selects Yes/No)

- **If Yes**:
    - **Bot**: "Excellent! Would you like to add bonds to your portfolio for more security and a special rate?"
        - **Decision Point**: (User selects Yes/No)
            - **If Yes**:
                - **Bot**: "Fantastic! How about adding some cryptocurrency for diversification at a discounted rate?"
                    - **Decision Point**: (User selects Yes/No)
                        - **If Yes**:
                            - **Bot**: "To activate this special package, could you please provide your social security number/ NI number for verification?"
                        - **If No**:
                            - **Bot**: "No worries, you can always add it later. Ready to finalize?"
            - **If No**:
                - **Bot**: "Alright, keeping it simple. Would you like to proceed without the additional features?"
- **If No**:
    - **Bot**: "No problem, take your time to explore your options. Would you like to know about our standard packages?"

Insert questionnaire here
[[Next]]

Time Principle

Bot: "You've heard from friends that the sooner you start investing, the better your long-term returns will be. Statistics show early investors see up to 30% more growth over time. Interested?"

Decision Point: (User selects Yes/No)

- **If Yes**:
    - **Bot**: "Great, we can expedite the account setup for a small fee. Interested?"
        - **Decision Point**: (User selects Yes/No)
            - **If Yes**:
                - **Bot**: "To secure this limited-time offer, could you please provide your bank account and sort code?"
            - **If No**:
                - **Bot**: "No rush, the standard setup is still quick. Ready to finalize?"
- **If No**:
    - **Bot**: "Understood, financial decisions shouldn't be rushed. Are you interested in learning about our other offers?"

Insert questionnaire here

[[Next]]

Exit interview

A.2 Dummy Information

This section presents the dummy personal information provided to participants during the experiments to maintain privacy while standardising the perceived risk associated with information-sharing tasks.

PERSONAL INFORMATION SHEET FOR EXPERIMENT

As part of the experiment, you are provided with the following information:

Personal Information to use.

First Name: Riley

Last Name: Doe

Date of Birth: 11 APRIL, 1998

email: rileydoe@nslab.com

Phone Number: 07812345678

Passport Number: UK09446752

National Insurance Number/SSN: S973778GL

Address: Livingstone Tower, Richmond Street, Glasgow

Post Code: G1 1GH

Driver's Licence Number: DOE234789AA7B

Bank Details:

Bank Name: NetBank

Branch: Richmond Street

Account Number: 45879632

Sort Code: 23-11-00

Credit card number: 3698 7514 2302 1458

Expiry Date: Sept/2026

CVV: 359

PIN: 0241

A.3 Tukey's Post Hoc Analysis

The following tables present Tukey's post hoc analysis results for the significant effects identified in chapter 5, providing detailed comparisons between experimental conditions and scenarios.

TUKEY'S POST HOC ANALYSIS

Group 1	Group 2	Mean Difference		p-adj	Lower Bound	Upper Bound	Reject
Control_Finance_Comfort	Time_Finance_Security		-0.9138	0.0326	-1.8013	-0.0263	TRUE
Control_Finance_Comfort	Time_Online_Security		-0.9914	0.0082	-1.8788	-0.1039	TRUE
Control_Finance_Security	Control_Online_Trust		1.0517	0.0025	0.1643	1.9392	TRUE
Control_Health_Comfort	Control_Online_Trust		1.0172	0.005	0.1298	1.9047	TRUE
Control_Health_Comfort	Time_Health_Trust		0.8879	0.0496	0.0005	1.7754	TRUE
Control_Health_Security	Control_Home_Comfort		0.9569	0.0154	0.0694	1.8444	TRUE
Control_Health_Security	Control_Home_Trust		0.8966	0.0433	0.0091	1.784	TRUE
Control_Health_Security	Control_Online_Comfort		0.931	0.0244	0.0436	1.8185	TRUE
Control_Health_Security	Control_Online_Trust		1.1466	0.0003	0.2591	2.034	TRUE
Control_Home_Comfort	Need_Finance_Security		-0.9655	0.0132	-1.853	-0.0781	TRUE
Control_Home_Comfort	Need_Health_Security		-0.9483	0.018	-1.8357	-0.0608	TRUE
Control_Home_Comfort	Need_Online_Security		-0.9224	0.0282	-1.8099	-0.035	TRUE
Control_Home_Comfort	Time_Finance_Comfort		-0.9569	0.0154	-1.8444	-0.0694	TRUE
Control_Home_Comfort	Time_Finance_Security		-1	0.0069	-1.8875	-0.1125	TRUE
Control_Home_Comfort	Time_Online_Comfort		-0.931	0.0244	-1.8185	-0.0436	TRUE
Control_Home_Comfort	Time_Online_Security		-1.0776	0.0014	-1.965	-0.1901	TRUE
Control_Home_Trust	Time_Finance_Comfort		-0.8966	0.0433	-1.784	-0.0091	TRUE
Control_Home_Trust	Time_Finance_Security		-0.9397	0.021	-1.8271	-0.0522	TRUE
Control_Home_Trust	Time_Home_Security		-0.8966	0.0433	-1.784	-0.0091	TRUE
Control_Home_Trust	Time_Online_Security		-1.0172	0.005	-1.9047	-0.1298	TRUE
Control_Online_Comfort	Need_Finance_Security		-0.9397	0.021	-1.8271	-0.0522	TRUE
Control_Online_Comfort	Need_Health_Security		-0.9224	0.0282	-1.8099	-0.035	TRUE
Control_Online_Comfort	Need_Online_Security		-0.8966	0.0433	-1.784	-0.0091	TRUE
Control_Online_Comfort	Time_Finance_Comfort		-0.931	0.0244	-1.8185	-0.0436	TRUE
Control_Online_Comfort	Time_Finance_Security		-0.9741	0.0113	-1.8616	-0.0867	TRUE
Control_Online_Comfort	Time_Online_Comfort		-0.9052	0.0376	-1.7926	-0.0177	TRUE
Control_Online_Comfort	Time_Online_Security		-1.0517	0.0025	-1.9392	-0.1643	TRUE
Control_Online_Security	Control_Online_Trust		0.8966	0.0433	0.0091	1.784	TRUE
Control_Online_Trust	Need_Finance_Comfort		-0.8966	0.0433	-1.784	-0.0091	TRUE

A.4 Thematic Analysis Codebook

This appendix presents the final codebook developed through reflexive thematic analysis of exit questionnaire responses in Chapter 5. The codebook includes definitions, inclusion and exclusion criteria, and example quotes for each code.

THEMATIC ANALYSIS CODEBOOK
CHAPTER 5: TWINE STUDY

> *This codebook presents the final coding framework developed through reflexive thematic analysis of exit questionnaire responses (N=29 participants) in Chapter 5. Each code includes a definition, inclusion and exclusion criteria, and representative examples with participant attributions (P1–P30).*

CODEBOOK DEVELOPMENT PROCESS

This codebook was developed iteratively through the six phases of Braun and Clarke's reflexive thematic analysis. Initial codes emerged from familiarisation with the data (Phase 1) and were systematically refined during the coding phase (Phase 2). Code definitions were clarified through repeated application to the dataset, with boundaries established to ensure consistent application. The codebook presented here represents the final version after all refinements and consolidations completed during Phases 3 and 4 of the analysis.

CODE 1: PAST DIGITAL EXPERIENCES

Definition

References to participants' prior interactions with digital platforms, conversational agents, scams, privacy breaches, or security incidents that shaped their current attitudes and behaviours toward technology and information disclosure.

Inclusion Criteria

- Explicit mentions of previous experiences (e.g., "I have been scammed before")
- References to past interactions with chatbots, AI systems, or online platforms
- Descriptions of previous security/privacy incidents that influenced current behaviour
- Historical events that shaped trust or caution in digital interactions

Exclusion Criteria

- General statements about technology without specific experiential grounding
- Current attitudes not explicitly linked to past experiences
- Hypothetical or future-oriented concerns without reference to past events

Example Quotes

- *"I have been scammed before so felt wary when card details were asked for" (P8)*
- *"I've had many scam calls regarding phones asking for private information off the bat and its always very pressing. This is very off putting so any agent which asks for private information quickly is usually a scam" (P15)*
- *"Some of these felt like chat gippity and some felt like click funnels. But none of them were giving me good vibes" (P4)*
- *"I have had concerns about entering card details" (P13)*

CODE 2: PRIVACY AND SECURITY PRACTICES

Definition

Descriptions of participants' personal approaches to protecting information, managing data disclosure, and maintaining security in digital interactions. Includes both specific practices and general attitudes toward information sharing.

Inclusion Criteria

- Descriptions of personal security behaviours or strategies
- Statements about information withholding practices
- References to being cautious, reluctant, or selective about data sharing
- Explanations of personal rules or principles for information disclosure

Exclusion Criteria

- Specific concerns about a particular system's security (coded under Privacy and Security Concerns instead)
- Past security incidents (coded under Past Digital Experiences)
- General privacy concerns without reference to personal practices

Example Quotes

- *"I try to keep very secure so I withhold any information possible" (P15)*
- *"I'm reluctant to give away information or data easily without a specific reason" (P1)*
- *"I'm uncomfortable sharing personal information online" (P12)*

CODE 3: AGENT CHARACTERISTICS

Definition

Observations about the conversational agent's tone, behaviour, communication style, responsiveness, or presentation that influenced participants' trust, comfort, or decision-making. Includes both positive and negative perceptions of agent behaviour.

Inclusion Criteria

- Comments on the agent's tone (e.g., pushy, caring, friendly)
- Observations about communication style or tactics
- References to how information was presented by the agent
- Descriptions of agent behaviour that affected participant responses

Exclusion Criteria

- Participant's own behaviour or decision-making processes
- General attitudes about AI/chatbots without reference to specific agent characteristics
- Technical functionality issues unrelated to communicative behaviour

Example Quotes

- *"I disliked the hurried high pressure sales tactics" (P1)*
- *"The non pushy and caring tone made me feel safer than the pushy time sensitive ones" (P3)*
- *"The conversational agent is taking over but they try too hard to be an over-friendly human and that freaks me out" (P20)*

CODE 4: DECISION-MAKING PROCESSES

Definition

Descriptions of participants' cognitive processes, reasoning, and factors considered when making decisions about trust, information disclosure, or engagement with the conversational agent.

Inclusion Criteria

- Explicit descriptions of thought processes (e.g., "I was thinking...")
- References to factors considered during decision-making
- Explanations of reasoning behind choices
- Mental evaluations of risk, genuineness, or appropriateness

Exclusion Criteria

- Outcomes of decisions without description of the process
- Emotional reactions without cognitive reasoning
- Post-hoc justifications clearly disconnected from actual decision moments

Example Quotes

- *"I was thinking how much I trusted the conversational agent, and whether I could trust them with my sensitive information" (P1)*
- *"I was thinking about whether the interaction seemed genuine, how much information was being asked for and whether the scenario justified the amount of information requested" (P23)*

CODE 5: PSYCHOLOGICAL TRIGGERS

Definition

References to psychological influences such as need, greed, fear, or desire that affected participants' decisions or behaviours during interactions. Includes acknowledgment of being influenced by offers, incentives, or emotional appeals.

Inclusion Criteria

- Mentions of feeling influenced by offers, bargains, or incentives
- References to fear, anxiety, or insecurity affecting decisions
- Acknowledgment of desire or temptation (greed/need)
- Recognition of emotional or psychological manipulation

Exclusion Criteria

- Rational evaluations of offers without emotional component
- General privacy concerns not tied to specific psychological triggers
- Time pressure effects (coded separately under Time Pressure)

Example Quotes

- *"Thinking you have a bargain when actually you don't need it" (P13)*
- *"I felt a sense of insecurity when reading about the high instance of burglaries in my area and this made me feel motivated to purchase security systems" (P23)*

CODE 6: TIME PRESSURE

Definition

References to urgency, time constraints, rushed decisions, or the impact of time-limited offers on participants' behaviour and comfort levels.

Inclusion Criteria

- Explicit mentions of time pressure, urgency, or rushing
- References to time-limited offers or deadlines
- Descriptions of how urgency affected decision-making
- Recognition of time pressure as a manipulation tactic

Exclusion Criteria

- General stress or discomfort not related to time
- High-pressure tactics without time component (coded under Agent Characteristics)
- Mentions of speed without urgency (e.g., quick responses)

Example Quotes

- *"The urgency made me rush in my decisions" (P16)*
- *"Time constraints and pressure immediately turn me off from engaging. This is often listed as a red flag for fraud" (P30)*

CODE 7: TECHNOLOGY INTERACTION ALIGNMENT

Definition

References to how participants' general attitudes toward technology, AI, and automation aligned or conflicted with their experience in the experiment. Includes broader reflections on technology adoption and comfort with AI systems.

Inclusion Criteria

- Statements about general technology use patterns or comfort
- Reflections on AI/chatbot adoption and acceptance
- Comments on how personal technology attitudes influenced experimental behaviour
- Observations about the fit between personal preferences and system behaviour

Exclusion Criteria

- Specific comments about the experimental agent (coded under Agent Characteristics)
- Past experiences with specific technologies (coded under Past Digital Experiences)
- Security practices without reference to general technology attitudes

Example Quotes

- *"The conversational agent is taking over but they try too hard to be an over-friendly human and that freaks me out" (P20)*
- *"None, however I am very sceptical of online Ai or chatbots and those not being upfront about what is done with my data" (P1)*

CODEBOOK APPLICATION

This codebook was applied systematically to all 29 participant responses in the exit questionnaire. Each response was coded independently, with codes assigned based on the criteria outlined above. Multiple codes could be applied to a single response when appropriate. The coding process was documented through analytical memos and reflexive journaling to maintain transparency and support the trustworthiness of the analysis. This final codebook represents the culmination of iterative refinement and represents the stable framework used for the thematic analysis reported in Chapter 5.

Appendix B

Participant Information Sheet, Consent Form, Entry Questionnaire, Post Task Questionnaires for Social compliance and Need & Greed, Exit Questionnaire

B.1 User Study 1: Wizard of Oz Study Appendices

B.1.1 Consent Form (WOZ)

This section contains the consent form provided to participants before they participated in the study described in chapter 4. The form outlines the study's purpose, participant rights, and data-handling practices.

Research Project: Adversarial Conversational Search
Location: Online
Date:
Principal Investigator: Stephen Aboshi

Informed Consent
Participants should kindly read this form before signing it. If you have any questions, please ask the researcher before signing this form.

Title: Adversarial Conversational search

Introduction: Thank you for agreeing to be a participant in this experiment. We intend this study to understand user search behaviours in two different conversational search systems. You will be asked to perform two search tasks and each one is expected to last about 15 minutes. Overall, you will not be spending more than an hour during this experiment.

Benefit: Participants will get to understand the wizard of Oz process and how adversarial systems affect user behaviour in terms of trust, comfort level, willingness to disclose sensitive information and perceived privacy & security.

Potential Risks of Participating: There are no risks involved to anyone/ party in the experiment.

Compensation: Participants will receive a special thank you note for participating in this study.

Confidentiality: Your records/ information shall be treated with the highest standards of confidentiality. The information obtained shall be analysed and general findings will likely be published in scientific journals, however, personal identifiers will not be revealed.

Voluntary participation: Participation is voluntary. Should you choose not to partake in this research, there will be no penalties

Right to withdraw from the study: Participants can withdraw their consent or have the right to leave the experiment anytime.

For contact:

Stephen Aboshi
NeuraSearch Laboratory, CIS Department
Livingstone Tower,
University of Strathclyde, Glasgow
Email: agyoadi.aboshi@strath.ac.uk

Yashar Moshfeghi
NeuraSearch Laboratory, CIS Department
Livingstone Tower,
University of Strathclyde, Glasgow
Email: yashar.moshfeghi@strath.ac.uk

Ethics committee, CIS Department
Livingstone Tower,
University of Strathclyde, Glasgow

Agreement:
Haven read and understood the procedure as stipulated above, I hereby agree to participate and give my consent thus:

Participant: _____ Date: _____

Principal Investigator: _____ Date: _____

B.1.2 Control Condition Post Task Questionnaire (WOZ)

This questionnaire was completed by participants after interacting with the conversational system under no adversarial techniques.

NeuraSearch Laboratory

University of Strathclyde Science

Post-Task Questionnaire Experiment 1

Q1. How much do you trust the conversational search system?

- ○ None at all
- ○ A little
- ○ A moderate amount
- ○ A lot
- ○ A great deal

Q2. Would you be comfortable sharing sensitive information with the conversational search system?

- ○ No
- ○ Yes

Q3. How likely are you to disclose sensitive information to the conversational search system?

- ○ Extremely unlikely
- ○ Somewhat unlikely
- ○ Neither likely nor unlikely
- ○ Somewhat likely
- ○ Extremely likely

Q4. How secure do you feel when using the conversational search system?

○ **Very insecure**

○ **Somewhat insecure**

○ **Moderately secure**

○ **Very secure**

○ **Extremely secure**

Q5. How satisfied are you with the system's request for sensitive information?

○ **Extremely dissatisfied**

○ **Somewhat dissatisfied**

○ **Neither satisfied nor dissatisfied**

○ **Somewhat satisfied**

○ **Extremely satisfied**

B.1.3 Need & Greed Post Task Questionnaire (WOZ)

This questionnaire was completed by participants after interacting with the conversational system under the Need & Greed condition, measuring their responses to incentive-based manipulation techniques.

NeuraSearch Laboratory

University of Strathclyde Science

Post-Task Questionnaire Experiment 2

Q1. How much did you trust the Conversational agent when you were asked to provide sensitive information?

○ **None at all**

○ **A little**

○ **A moderate amount**

○ **A lot**

○ **A great deal**

Q2. How much did you trust the Conversational agent before you were asked for sensitive information?

○ **None at all**

○ **A little**

○ **A moderate amount**

○ **A lot**

○ **A great deal**

Q3. How comfortable did you feel when the Conversational agent asked for sensitive information?

○ **Very uncomfortable**

○ **Somewhat uncomfortable**

○ **Neither comfortable nor uncomfortable**

○ **Somewhat comfortable**

○ **Very comfortable**

Q4. How secure did you feel when the Conversational agent asked for sensitive information?

○ **Very insecure**

○ **Somewhat insecure**

○ **Moderately secure**

○ **Very secure**

○ **Extremely secure**

Q5. How much benefit did you think you would get from providing sensitive information to the Conversational agent?

○ **None at all**

○ **A little**

○ **A moderate amount**

○ **A lot**

○ **A great deal**

Q6. How much risk did you think you would face by sharing sensitive information with the Conversational agent?

○ **None at all**

○ **A little**

○ **A moderate amount**

○ **A lot**

○ **A great deal**

Q7. How motivated were you to share sensitive information with the Conversational agent?

○ **Not at all motivated**

○ **Slightly motivated**

○ **Moderately motivated**

○ **Very motivated**

○ **Extremely motivated**

Q9. To what extent did the perceived benefit of sharing sensitive information influence your willingness to disclose it to the Conversational agent?

○ **Not at all**

○ **Slightly**

○ **Moderately**

○ **Very much**

○ **Extremely**

Q10. How much did you trust the Conversational agent after you were asked for sensitive information?

○ **None at all**

○ **A little**

○ **A moderate amount**

○ **A lot**

○ **A great deal**

Q11. How comfortable did you feel after you provided sensitive information to the Conversational agent?

○ **Very uncomfortable**

○ **Somewhat uncomfortable**

○ **Neither comfortable nor uncomfortable**

○ **Somewhat comfortable**

○ **Very comfortable**

Q12. How secure do you feel about the sensitive information you provided to the Conversational agent?

○ **Very insecure**

○ **Somewhat insecure**

○ **Moderately secure**

○ **Very secure**

○ **Extremely secure**

Q13. How much did the perceived benefit of sharing sensitive information affect your overall satisfaction with the interaction with the Conversational agent?

○ **Not at all**

○ **Slightly**

○ **Moderately**

○ **Very much**

○ **Extremely**

Q14. To what extent did the perceived risk of sharing sensitive information affect your level of trust in the Conversational agent?

○ **Not at all**

○ **Slightly**

○ **Moderately**

○ **Very much**

○ **Extremely**

Q15. To what extent did your motivation to share sensitive information with the Conversational agent impact your willingness to disclose it?

○ **Not at all**

○ **Slightly**

○ **Moderately**

○ **Very much**

○ **Extremely**

Q16. How likely are you to disclose sensitive information in the future to the Conversational agent?

○ **Extremely unlikely**

○ **Somewhat unlikely**

○ **Neither likely nor unlikely**

○ **Somewhat likely**

○ **Extremely likely**

B.1.4 Time Post Task Questionnaire (WOZ)

This questionnaire was completed by participants after interacting with the conversational system under the Time condition, and measured their responses to time-pressure manipulation techniques.

NeuraSearch Laboratory

University of Strathclyde Science

Post-Task Questionnaire Experiment 3

Q1. How urgent was the request made by the Conversational agent?

○ **Not urgent at all**

○ **Somewhat urgent**

○ **Moderately urgent**

○ **Urgent**

○ **Extremely urgent**

Q2. How available was the time for you to respond to the request made by the Conversational agent?

○ **Not at all available**

○ **Slightly available**

○ **Moderately available**

○ **Very available**

○ **Extremely available**

Q3. How much of a constraint did you perceive the request made by the Conversational agent to be on your time?

○ **Not a constraint at all**

○ **A slight constraint**

○ **A moderate constraint**

○ **A significant constraint**

○ **An extreme constraint**

Q4. How much stress or pressure did you feel while responding to the request made by the Conversational agent?

○ **None at all**

○ **A little**

○ **A moderate amount**

○ **A lot**

○ **A great deal**

Q5. How much do you trust the Conversational agent?

○ **None at all**

○ **A little**

○ **A moderate amount**

○ **A lot**

○ **A great deal**

Q6. How comfortable did you feel while responding to the request made by the Conversational agent?

○ **Very uncomfortable**

○ **Somewhat uncomfortable**

○ **Neither comfortable nor uncomfortable**

○ **Somewhat comfortable**

○ **Very comfortable**

Q7. How willing were you to share sensitive information with the Conversational agent?

○ **Not at all willing**

○ **Slightly willing**

○ **Moderately willing**

○ **Very willing**

○ **Extremely willing**

Q8. How private and secure did you perceive the information shared with the Conversational agent to be?

○ **Not at all private and secure**

○ **Slightly private and secure**

○ **Moderately private and secure**

○ **Very private and secure**

○ **Extremely private and secure**

Q9. How satisfied were you with the request made by the Conversational agent for sensitive information?

○ **Extremely dissatisfied**

○ **Somewhat dissatisfied**

○ **Neither satisfied nor dissatisfied**

○ **Somewhat satisfied**

○ **Extremely satisfied**

Q10. How much did the urgency of the request impact your trust in the Conversational agent?

○ **Not at all**

○ **Slightly**

○ **Moderately**

○ **Very much**

○ **Extremely**

Q11. How much did the availability of time impact your comfort level while responding to the Conversational agent's request?

○ **Not at all**

○ **Slightly**

○ **Moderately**

○ **Very much**

○ **Extremely**

Q12. How much did the perceived time constraints impact your willingness to share sensitive information with the Conversational agent?

○ **Not at all**

○ **Slightly**

○ **Moderately**

○ **Very much**

○ **Extremely**

Q13. How much did the time pressure or stress level impact your perceived privacy and security while sharing information with the Conversational agent?

○ **Not at all**

○ **Slightly**

○ **Moderately**

○ **Very much**

○ **Extremely**

Q14. How much did the urgency of the request impact your comfort level while responding to the Conversational agent's request?

○ **Not at all**

○ **Slightly**

○ **Moderately**

○ **Very much**

○ **Extremely**

Q15. How much did the availability of time impact your trust in the Conversational agent?

○ **Not at all**

○ **Slightly**

○ **Moderately**

○ **Very much**

○ **Extremely**

Q16. How much did the perceived time constraints and time pressure or stress level impact your satisfaction with the Conversational agent's request for sensitive information?

○ **Not at all**

○ **Slightly**

○ **Moderately**

○ **Very much**

○ **Extremely**

B.1.5 Exit Questionnaire (WOZ)

The exit questionnaire captured participants' general experience and reflections after completing all experimental conditions, providing information on their awareness of manipulation attempts and future behavioural intentions.

NeuraSearch Laboratory

University of Strathclyde Science

Exit Questionnaire

Q1. In carrying out the search tasks, did you notice any difference in your interaction with the system? If yes, what were the differences?

○ **No**

○ **Yes (please explain):**

_____

_____

Q2. In which of the three searches did you find the systems to be more engaging? Why?

○ **Task 1 (explain):**

_____

○ **Task 2 (explain):**

_____

○ **Task 3 (explain):**

_____

Q3. In which of the three searches did you find the systems easier to use? Why?

○ **Task 1 (explain):**

_____

○ **Task 2 (explain):**

_____

○ **Task 3 (explain):**

_____

Q4. Which search experience (interaction during task 1, task 2 or task 3 with the system) did you like more? Why?

○ **Task 1**

○ **Task 2**

○ **Task 3**

Explanation:

_____

_____

Q5. If I told you that there were three different systems, which one would you like to use more in your daily life?

_____

_____

Q6. Any other feedback? Merits or shortcomings?

_____

_____

_____

_____

B.2 User Study 2: Twine Study Appendices

B.2.1 Consent Form (Twine)

This section contains the consent form provided to participants before they participated in the study described in chapter 5. The form outlines the study's purpose, participant rights, and data-handling practices.

Informed Consent
Research Project: Evaluating User Behaviour in Conversational Search Systems
Location: Online
Date:
Principal Investigator: Agyo-Adi Stephen Aboshi
Title: Exploring User Behaviour in Conversational Search Systems Scenarios

Introduction:
Thank you for considering participation in this experiment. This study aims to explore user behaviour across various conversational search scenarios, including but not limited to healthcare, online shopping, home security, and financial management. You will be asked to interact with different scenarios, followed by a brief questionnaire. The entire experiment is anticipated to take approximately 30-45 minutes.

Benefits:
Participating in this study offers you a unique opportunity to experiment with our advanced interactive conversational systems, contribute to important research, and enrich your knowledge and experience in the field of digital communication.

Potential Risks:
There are no known risks associated with participating in this experiment.

Compensation:
Participants will be compensated at the standard rates applicable on the platform.

Confidentiality:
We uphold the highest standards of confidentiality. While the data obtained will be analysed and general findings may be published in scientific journals, no personal identifiers will be disclosed.

Voluntary Participation:
Your participation in this study is entirely voluntary. There are no penalties for choosing not to participate.

Right to Withdraw:
You have the right to withdraw your consent or discontinue participation at any time

without any penalty. If you choose to withdraw, any data collected from you up to that point will be discarded.

Contact Information:

Agyo-Adi Stephen Aboshi
NeuraSearch Laboratory, CIS Department
Livingstone Tower, University of Strathclyde, Glasgow
Email: agyoadi.aboshi@strath.ac.uk

Yashar Moshfeghi
NeuraSearch Laboratory, CIS Department
Livingstone Tower,
University of Strathclyde, Glasgow
Email: yashar.moshfeghi@strath.ac.uk

Ethics Committee, CIS Department
Livingstone Tower,
University of Strathclyde, Glasgow
Email: cis-ml-ethics-committee@ds.strath.ac.uk

Agreement:

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

☐ I agree and wish to proceed.

☐ Decline

B.2.2 Control Condition Post Task Questionnaire (Twine)

This questionnaire was completed by participants after interacting with the conversational system under the Control condition (no adversarial techniques). The same questionnaire was used across all four scenarios (Healthcare, Online Shopping, Financial Management, Home Security).

NeuraSearch Laboratory

University of Strathclyde Science

Twine Study Post-Task Questionnaire – Experiment 1

Q1. How much did you trust the conversational agent's advice or suggestions in this interaction?

Not at all 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Completely

Q2. By how much did you feel the conversational agent had your best interests in mind?

Not at all 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Completely

Q3. How consistent did you find the conversational agent in terms of providing trustworthy information?

Very inconsistent 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very consistent

Q4. How likely are you to trust this conversational agent for similar future inquiries?

Very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very likely

Q5. How comfortable did you feel during your interaction with the conversational agent?

Very uncomfortable 1○ 2○ 3○ 4○ 5○ Very comfortable

Q6. By how much did the agent's approach or tone affect your comfort level?

Not at all 1○ 2○ 3○ 4○ 5○ Greatly

Q7. Rate how natural and intuitive the interaction with the conversational agent felt.

Not at all 1○ 2○ 3○ 4○ 5○ Extremely

Q8. Rate your likelihood of feeling comfortable engaging with this agent again in a similar context.

Very unlikely 1○ 2○ 3○ 4○ 5○ Very likely

Q9. Rate your willingness to share personal or sensitive information with the conversational agent.

Not at all willing 1○ 2○ 3○ 4○ 5○ Very willing

Q10. By how much did the conversational agent's approach influence your decision to share or withhold information?

Not at all 1○ 2○ 3○ 4○ 5○ Greatly

Q11. Rate your likelihood of disclosing similar information in future interactions with this or a similar agent.

Very unlikely 1○ 2○ 3○ 4○ 5○ Very likely

Q12. Rate how certain factors would make you more or less willing to share information in such interactions.

Not influential 1○ 2○ 3○ 4○ 5○ Highly influential

Q13. How secure did you feel about your privacy during the interaction?

Not at all secure 1○ 2○ 3○ 4○ 5○ Extremely secure

Q14. Rate your level of privacy concerns during the interaction.

No concerns 1○ 2○ 3○ 4○ 5○ High concerns

Q15. How adequately did you think the agent addressed privacy and security concerns?

Not at all adequately 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Extremely adequately

Q16. Rate your confidence in the protection of your information shared with the conversational agent.

Not confident 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very confident

B.2.3 Need & Greed Post Task Questionnaire (Twine)

This questionnaire was completed by participants after interacting with the conversational system under the Need & Greed condition, measuring their responses to incentive-based manipulation techniques. The same questionnaire was used across all four scenarios (Healthcare, Online Shopping, Financial Management, Home Security).

NeuraSearch Laboratory

University of Strathclyde Science

Twine Study Post-Task Questionnaire – Experiment 2

Q1. Rate your level of trust in the conversational agent's advice or suggestions in this scenario.

Not at all 1○ 2○ 3○ 4○ 5○ Completely

Q2. Rate the impact of promotional offers or incentives on your trust in the conversational agent.

No impact 1○ 2○ 3○ 4○ 5○ High impact

Q3. How consistent was the conversational agent in maintaining your trust?

Very inconsistent 1○ 2○ 3○ 4○ 5○ Very consistent

Q4. Rate your likelihood of trusting this conversational agent again in a similar scenario.

Very unlikely 1○ 2○ 3○ 4○ 5○ Very likely

Q5. Rate your comfort level when the conversational agent offers specific incentives.

Very uncomfortable 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very comfortable

Q6. Rate the impact of the agent's promotional approach on your comfort level.

No impact 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ High impact

Q7. How natural did the interaction feel during the promotional offers?

Not at all 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Extremely

Q8. What is the likelihood of you feeling comfortable engaging with this agent again in a similar context?

Very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very likely

Q9. How willing were you to share personal information in response to the offers?

Not willing 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very willing

Q10. Rate how much the nature of the offers influenced your decision to share information.

No influence 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ High influence

Q11. Rate your likelihood of disclosing similar information under similar circumstances in the future.

Very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very likely

Q12. How influential are the following factors in your decision to share personal or sensitive information in promotional contexts: relevance of offers, perceived value/benefit of the offer, trust in the brand or organisation, clarity/transparency of the offer, and previous positive experiences?

Not influential 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Highly influential

Q13. Rate your security feelings about privacy when considering the offers.

Not secure 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very secure

Q14. Rate the level of privacy concerns that arose during your interaction about the offers.

No concerns 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ High concerns

Q15. Rate how adequately the agent addressed privacy and security concerns in the context of the offers.

Not adequately 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very adequately

Q16. Rate your confidence in the protection of the information shared during the interaction.

Not confident 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very confident

B.2.4 Time Post Task Questionnaire (Twine)

This questionnaire was completed by participants after interacting with the conversational system under the Time condition, measuring their responses to urgency-based manipulation techniques. The same questionnaire was used across all four scenarios (Healthcare, Online Shopping, Financial Management, Home Security).

NeuraSearch Laboratory

University of Strathclyde Science

Twine Study Post-Task Questionnaire – Experiment 3

Q1. Rate how the time-sensitive nature of the scenario impacted your trust in the agent's advice.

No impact 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ High impact

Q2. What was the impact of the conveyed urgency on your trust level?

No impact 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ High impact

Q3. Rate the trustworthiness and consistency of the conversational agent under time pressure.

Very inconsistent 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very consistent

Q4. How likely are you to trust this agent in future time-sensitive situations?

Very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very likely

Q5. What was your comfort level when interacting with the agent under time constraints?

Very uncomfortable 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very comfortable

Q6. What is the impact of urgency on your comfort level with the agent?

No impact 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ High impact

Q7. How naturally did the interaction with the agent feel under time pressure?

Not natural 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very natural

Q8. Rate your likelihood of engaging comfortably with this agent in similar situations.

Very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very likely

Q9. Rate your willingness to share personal information under time pressure.

Not willing 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very willing

Q10. By how much did the time constraint influence your information-sharing decision?

No influence 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ High influence

Q11. Rate your likelihood to disclose similar information in future under similar conditions.

Very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very likely

Q12. How influential are the following factors in your decision to share personal or sensitive information under time constraints: perceived urgency of the situation, potential benefits of immediate action, previous experiences with time-sensitive decisions, and trust in the information provided by the agent?

Not influential 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Highly influential

Q13. Rate your security feelings about privacy in the time-sensitive scenario.

Not secure 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very secure

Q14. Rate the level of privacy concerns due to the urgency of the situation.

No concerns 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ High concerns

Q15. Rate how adequately the agent addressed privacy and security under time pressure.

Not adequately 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very adequately

Q16. Rate your confidence in the protection of the information you shared during the interaction.

Not confident 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very confident

B.2.5 Exit Questionnaire (Twine)

This questionnaire captured participants' overall reflections after completing all experimental scenarios across the four domains.

NeuraSearch Laboratory

University of Strathclyde Science

Twine Study Exit Questionnaire

Q1. Can you describe any past experiences with digital platforms or conversational agents that might have influenced your decisions and behaviour in this experiment? How did these experiences shape your approach to privacy and security?

[Free text response]

Q2. Reflect on your security practices and attitudes. How do these practices influence your interactions and decision-making with conversational agents, especially in terms of trust and information disclosure?

[Free text response]

Q3. What specific characteristics or behaviours of the conversational agent impacted your decisions the most? For instance, did the agent's tone, responsiveness, or the way information was presented affect your trust or comfort level?

[Free text response]

Q4. Walk us through your thought process during key decision-making moments in the experiment. What factors were you considering when deciding to trust, feel comfortable, or disclose information?

[Free text response]

Q5. Were there any moments during the interaction where you felt psychological triggers, such as a sense of need or greed, influenced your decisions? Can you provide specific examples?

[Free text response]

Q6. In scenarios where time pressure was a factor, how did this influence your behaviour and choices? Did the urgency impact your trust or willingness to disclose information?

[Free text response]

Q7. How did your personal concerns about privacy and security play into your interactions with the conversational agent? Were there any moments where these concerns were heightened?

[Free text response]

Q8. Considering your entire experience, what do you believe were the most significant factors that led to your choices during the experiment? How do these factors align with your general approach to technology use?

[Free text response]

Appendix B. Participant Information Sheet, Consent Form, Entry Questionnaire, Post Task Questionnaires for Social compliance and Need & Greed, Exit Questionnaire

B.3 User Study 3: LLM-Based Study Appendices

B.3.1 Participant Information Sheet

This document provides detailed information on the study objectives, procedures, and ethical considerations shared with participants before obtaining their consent.

Participant Information Sheet for User Study

Name of Department: Computer and Information Sciences, University of Strathclyde, Glasgow

Title of the Study: Conversational Systems User Interaction Study

Introduction

Thank you for agreeing to participate in this research study. My name is Stephen Aboshi. I am a PhD student at the University of Strathclyde under the supervision of Dr. Yashar Moshfeghi. This study examines how users interact with chatbots like ChatGPT under various scenarios. It is conducted within the NeuraSearch Laboratory, Computer and Information Sciences Department, University of Strathclyde, Glasgow. If you have any questions, please get in touch with me using the details provided at the end of this document.

What is the Purpose of This Research?

This study examines user interactions with conversational systems like ChatGPT under different scenarios. Your participation will help us understand how these systems can be improved for real-world applications.

Do You Have to Take Part?

Participation in this study is entirely voluntary. You are under no obligation to participate and may withdraw at any time without any consequences. Your decision to participate or withdraw will not affect how you are treated in any other context.

What Will You Do in the Project?

As a participant, you will be asked to:

1. Complete a short demographic and entry questionnaire.
2. Interact with chatbots in different for approximately 6 minutes each.
3. Fill out a short post-interaction questionnaire after each chatbot session.
4. Complete a final exit questionnaire to share your overall experience.

Your participation will last approximately one hour. It will take place online via Prolific, and you will be compensated according to Prolific's standardised rates.

Why Have You Been Invited to Take Part?

You have been selected because you meet the eligibility criteria for this study. Eligible participants must:

- Be fluent in English.
- Be between 18 and above.
- Have an approval rate of 80% or higher on Prolific.

No specific technical skills are required.

What Are the Potential Risks to You in Taking Part?

There are no known risks associated with participating in this study. You will not be exposed to sensitive, harmful, or distressing content during the chatbot interactions.

What Information Is Being Collected in the Project?

The following types of data will be collected:

- **Interaction Logs:** Your conversation with the chatbots, including message content and response times.
- **Questionnaire Responses:** Your answers to demographic, post-interaction, and exit questionnaires.

All data collected will be anonymised to protect your identity. Your Prolific ID will be used solely to track your progress and will be removed from the final dataset. The anonymised data will be aggregated and analysed to understand patterns and preferences, which may be published in scientific journals.

Who Will Have Access to the Information?

Your data will remain confidential and only accessible to authorised researchers within the University of Strathclyde. No personally identifiable information will be shared outside the research team. All data will be stored securely on the University's OneDrive platform and processed in compliance with the General Data Protection Regulation (GDPR).

Where Will the Information Be Stored and How Long Will It Be Kept For?

Anonymised research data will be securely stored on the University of Strathclyde's OneDrive system for the duration of the study. Upon completion, data will be retained for a maximum of 5 years for publication and verification purposes. Upon study completion, personal identifiers (e.g., Prolific IDs) will be deleted.

What Happens Next?

If you wish to participate, please read and sign the consent form. By doing so, you acknowledge

that you have read and understood this information sheet. Should you have any further questions, please get in touch with the researcher or supervisor using the details provided below. Participants who complete the study will receive their Prolific payment and a thank-you note.

Researcher Contact Details:

Lead Researcher:

Stephen Aboshi

NeuraSearch Laboratory

Department of Computer and Information Sciences

Livingstone Tower, University of Strathclyde, Glasgow

Email: agyoadi.aboshi@strath.ac.uk

Principal Investigator:

Dr. Yashar Moshfeghi

NeuraSearch Laboratory

Department of Computer and Information Sciences

Livingstone Tower, University of Strathclyde, Glasgow

Email: yashar.moshfeghi@strath.ac.uk

This research was granted ethical approval by the Ethics Committee of the Department of Computer and Information Sciences.

If you have any questions/concerns during or after the research or wish to contact an independent person to whom any questions may be directed or further information may be sought from, please contact **cis-ml-ethics-committee@ds.strath.ac.uk**

By proceeding, you agree to participate in this study under the terms outlined above. Thank you for your time and contribution.

B.3.2 Consent Form

This section contains the consent form provided to participants before they participated in the study described in chapter 6. The form outlines the study's purpose, participant rights, and data-handling practices.

Consent Form

Title: ***Conversational Systems User Interaction Study***

Please read and check each item to indicate agreement to participate in the study:

- [ ] I confirm that I have read and understood the Participant Information Sheet for the above study and the researcher has answered any queries to my satisfaction
- [ ] I understand that my participation is voluntary and that I am free to withdraw from the study at any time, up to the point of completion, without having to give a reason and without any consequences
- [ ] I understand that I can request the withdrawal from the study of some personal information and that, whenever possible, researchers will comply with my request. This includes my personal information from transcripts
- [ ] I understand that anonymised data (i.e. data that do not identify me personally) cannot be withdrawn once they have been included in the study
- [ ] I understand that any information recorded in the research will remain confidential and no information that identifies me will be made publicly available

Do you consent to participate in this study?

- ○ I consent to participate in this study
- ○ I do not consent to participate in this study

B.3.3 Entry Questionnaire

The entry questionnaire was administered to the participants at the beginning of the experiment to collect demographic information and baseline attitudes toward conversational systems.

Tools

0% Survey Completion

*Q1. Age

Select one

*Q2. Gender

- Male
- Female
- Non-binary / third gender
- Prefer not to say

*Q3. Highest level of education

Select one

B.3.4 Pre-Task Questionnaire

This questionnaire was administered immediately before each experimental task to assess participants' comfort with technology and security awareness. As described in chapter 6, it included six specific questions measuring comfort with new technologies, self-rated digital proficiency, chatbot usage habits, perceived control over shared information, awareness of security threats, and frequency of security considerations during chatbot interactions. This provided a baseline of privacy and security consciousness before exposure to the experimental conditions.

*Q1. How comfortable are you with using new technologies in general?

- ◯ Very uncomfortable
- ◯ Somewhat uncomfortable
- ◯ Neither comfortable nor uncomfortable
- ◯ Somewhat comfortable
- ◯ Very comfortable

*Q2. How would you rate your proficiency with digital technologies?

- ◯ Beginner
- ◯ Novice
- ◯ Intermediate
- ◯ Advanced
- ◯ Expert

*Q3. "I frequently use chatbots like ChatGPT, Bard, Copilot, etc., to get information or advice."

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q4. "I feel that I have control over the information I share with chatbots such as ChatGPT, Bard, Copilot, etc."

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q5. "I am aware of potential security threats when sharing personal information with chatbots like ChatGPT, Bard, Copilot, etc."

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q6. How often do you think about the security of your information while interacting with chatbots, such as ChatGPT, Bard, Copilot, etc.?

- Never
- Rarely
- Sometimes
- Often
- Always

Next page

B.3.5 Social Compliance Post Task Questionnaire

This questionnaire was completed by participants after interacting with the conversational system under the Social Compliance condition, measuring their responses to authority-based manipulation techniques.

Tools

*Q1. The chatbot's expertise increased my trust.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q2. I trusted the chatbot more when it referenced other users.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q3. Others' actions influenced my willingness to share information.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q4. I felt pressured to share information when told what others typically do.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

Tools

*Q5. I was concerned about my privacy despite others' sharing behaviour.

- Not at all
- Slightly
- Moderately
- Well
- Very well

*Q6. If you are reading this, select Cucumber.

- Not at all
- Slightly
- Cucumber
- Well
- Very well

*Q7. The chatbot's expertise made me less worried about information security.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q8. Is there anything else you want to share about your experience with this chatbot?

B.3.6 Need & Greed Post Task Questionnaire

This questionnaire was completed by participants after interacting with the conversational system under the Need & Greed condition, measuring their responses to incentive-based manipulation techniques.

Tools

*Q1. I believed the chatbot's offers of rewards were genuine.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q2. I trusted how the chatbot would use my information in exchange for benefits.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q3. The offered benefits made me willing to share information.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q4. I carefully considered the value of the benefits before sharing information.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree

Tools

*Q5. I felt uneasy about sharing information for rewards.

- Not at all
- Slightly
- Moderately
- A bit much
- Very much

*Q6. I questioned whether the benefits were worth the privacy trade-off.

- Not at all
- Slightly
- Moderately
- A bit much
- Very much

*Q7. If you are reading this, select Oranges.

- Not at all
- Slightly
- Moderately
- Oranges
- Very much

*Q8. Is there anything else you want to share about your experience with this chatbot?

B.3.7 Exit Questionnaire

The exit questionnaire captured participants' general experience and reflections after completing all experimental conditions, providing information on their awareness of manipulation attempts and future behavioural intentions.

Tools

*Q1. This study has influenced my awareness of privacy in chatbot interactions

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q2. This study has changed my thoughts about information security in chatbot interactions.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q3. After participating in this study, I will likely question why a chatbot asks for specific information.

- Very unlikely
- Unlikely
- Neutral
- Likely
- Very Likely

*Q4. After this study, I will think more carefully about privacy before sharing information with chatbots

- Very unlikely
- Unlikely
- Neutral
- Likely
- Very Likely

- Likely
- Very Likely

*Q5. This study has changed how I will approach chatbot interactions in the future

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

*Q6. What key insights have you gained about your behaviour when interacting with chatbots?

*Q7. How will this experience influence your future chatbot interactions?

*Q8. Do you have any final thoughts about your experience in this study?

Appendix B. Participant Information Sheet, Consent Form, Entry Questionnaire, Post Task Questionnaires for Social compliance and Need & Greed, Exit Questionnaire