Banal Deception and Human-AI Ecosystems: A Study of People's Perceptions of LLM-generated Deceptive Behaviour

XIAO ZHAN*, King's College London, UK
YIFAN XU*, The University of Manchester, UK
NOURA ABDI, Liverpool John Moores University, UK
JOE COLLENETTE, The University of Chester, UK
STEFAN SARKADI†, King's College London, UK

Large language models (LLMs) can provide users with false, inaccurate, or misleading information, and we consider the output of this type of information as what Natale calls 'banal' deceptive behaviour. Here, we investigate people's perceptions of ChatGPT-generated deceptive behaviour and how this affects people's behaviour and trust. To do this, we use a mixed-methods approach comprising (i) an online survey with 220 participants and (ii) semi-structured interviews with 12 participants. Our results show that (i) the most common types of deceptive information encountered were over-simplifications and outdated information; (ii) humans' perceptions of trust and chat-worthiness of ChatGPT are impacted by 'banal' deceptive behaviour; (iii) the perceived responsibility for deception is influenced by education level and the perceived frequency of deceptive information; and (iv) users become more cautious after encountering deceptive information, but they come to trust the technology more when they identify advantages of using it. Our findings contribute to understanding human-AI interaction dynamics in the context of *Deceptive AI Ecosystems* and highlight the importance of user-centric approaches to mitigating the potential harms of deceptive AI technologies.

JAIR Track: AI and Society

JAIR Associate Editor: Toby Walsh

JAIR Reference Format:
Xiao Zhan, Yifan Xu, Noura Abdi, Joe Collenette, and Stefan Sarkadi. 2025. Banal Deception and Human-AI Ecosystems: A Study of People's Perceptions of LLM-generated Deceptive Behaviour. *Journal of Artificial Intelligence Research* 84, Article 11 (October 2025), 30 pages. DOI: 10.1613/jair.1.18724

1 Introduction

According to Sarkadi (2023b), a *'deceptive AI ecosystem'* represents more than just the technical aspects of developing deceptive AI technologies . This ecosystem encompasses the mechanisms through which the societal and evolutionary pressures influence human interaction with deceptive AI technologies at multiple levels of interaction, i.e. as individuals, groups, organizations, and societies. These multi-layered interactions create an ever-evolving informational feedback loop between hybrid societies where humans and machines communicate

*Equal contribution.
†Corresponding Author.

Authors' Contact Information: Xiao Zhan, ORCID: 0000-0003-1755-0976, xiao.zhan@kcl.ac.uk, King's College London, London, UK; Yifan Xu, ORCID: 0000-0003-2303-1531, yifan.xu@manchester.ac.uk, The University of Manchester, Manchester, UK; Noura Abdi, ORCID: 0000-0002-4613-6443, n.a.abdi@ljmu.ac.uk, Liverpool John Moores University, Liverpool, UK; Joe Collenette, ORCID: 0000-0001-6179-2038, j.collenette@chester.ac.uk, The University of Chester, Chester, UK; Stefan Sarkadi, ORCID: 0000-0003-3999-528X, stefan.sarkadi@kcl.ac.uk, King's College London, London, UK.

DOI: 10.1613/jair.1.18724

as agents and the emerging socio-economical regulatory norms, human and societal values, business decisions, power structures, communication about AI technologies, and market behaviour.

In the past years, ChatGPT[1] has emerged as a powerful conversational AI system that has captured the attention of researchers and users alike. This advanced generative AI system is designed to generate human-like responses to user queries, making it an attractive tool for a range of applications, including customer service (Patterson 2024), and education (Kohnke, Moorhouse, and Zou 2023; Yangyu Xiao and Zhi 2023). Despite ChatGPT's remarkable performance, its potential for generating deceptive information in its responses to user prompts should not be ignored. Moreover, erroneous information provided by ChatGPT can lead to undesirable outcomes, causing users to lose trust in the system and impede its adoption in contexts in which it might prove useful (Zhan, Xu, and Sarkadi 2023).

Analogous to the distinction between Strong and Weak AI, there are two types of deceptive AI technologies that can act as agents within Deceptive AI ecosystems (Sarkadi 2023b). The first type comes in the form of fully autonomous AI agents whose cognitive architecture allows them to do the same thing human minds can do, which in this case is deceiving in the same way humans do. The development of the first type of deceptive AI technology follows the process that Boden described as *'making computers that do the same thing minds can do...'* (Boden 2016). In the case of human-like deception, this would be to engage in the process defined as *'The intentional process of an agent (the deceiver) to make another agent (the target) believe something is true (or false) that the deceiver believes is false (or true), with the aim of achieving an ulterior goal or desire'* (Sarkadi 2021). For the deceiver, this human-like cognitive, or better said meta-cognitive, process, implies, along with belief-formation, deliberation, and models of causal communication abilities (Searle 1969), mentalisation capabilities, i.e. the ability to form and use a Theory of Mind of the target. Yet, in this paper, we will not look at a machine's ability to deliberately deceive.

This brings us to the problem of developing the second type of deceptive AI technologies, which has less to do with the cognitive capabilities of AI agents themselves and, instead, has more to do with how humans perceive AI behaviour in different contexts; i.e., the effect of AI behaviour on humans' perceptions (Zhan, Xu, and Sarkadi 2023; Masters, Smith, et al. 2021). This second type of deceptive AI technology is not capable of engaging in the process of deception on its own, whether intentionally or deliberately, because such technologies lack the necessary cognitive modules and architectures to enable them to understand the meaning (semantics) or consequences of their actions in various contexts. They also lack the ability to form and use Theory of Mind (Verma, Bhambri, and Kambhampati 2024), which, together with meta-cognition and reflection, is a necessary ingredient for deception (Sarkadi 2021). Yet, their behaviour is nevertheless deceptive due to the biases of their human users and the context in which humans are interacting with them. AI researchers and engineers who build the second type of deceptive AI technologies aim to optimise the effect of AI behaviour on humans. From Natale's deceptive media perspective, this would mean that they optimise for 'banal' deception (Natale 2021).

Banal deception is not a process that an AI agent cognitively engages in. In this case, the AI agent does not require an ulterior motive or goal, nor the necessary cognitive capabilities to reason, plan and act to cause a desired false belief in the mind of its target. Banal deception arises from the contextual background in which the human-machine interaction takes place. Designers of technology can set up this background context in such a way that they control for banal deception by playing into humans' cognitive biases. I.e. humans become susceptible to false beliefs because inaccurate, misleading, or false informational content is provided in a context mediated by a technology that humans interact with.

In this way, banal deception can be triggered by specific contexts that allow for the deceptive information to be presented in ways that exploit humans' cognitive biases by keeping human targets in the cognitively efficient System 1 thinking (adopting mental shortcuts in making decisions or reaching conclusions) and avoiding to

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

trigger them into employing System 2 thinking (e.g., deliberation, argumentation, critical thinking) (Tversky and Kahneman 1988; Tversky and Kahneman 1996). This effect is also observed in human-human deception, where humans are in a truth-default state, a mental state that treats all incoming information as truthful, from which they only exit if something in the context seems 'fishy' (Levine 2019).

In this paper, we are tackling the problem posed by Large Language Models (LLMs), which are successful drivers of 'banal' deception that fall into the second type of deceptive AI technology. LLMs play into the tendency of humans to anthropomorphise (Schneiderman and Muller 2023). This tendency is driven by the cognitive biases of System 1 thinking (Epley, Waytz, and Cacioppo 2007). Moreover, due to the dynamic responses of LLM chatbot systems and due to their adaptability to human prompts, this anthropomorphic bias can be exploited in humans with individual differences (Letheren et al. 2016). The linguistic context of the human-LLM interaction also helps with anthropomorphisation (Kopp, Baumgartner, and Kinkel 2023).

LLM-based chatbots, including notable examples such as OpenAI's ChatGPT, Claude[2], Microsoft's Bing Chat[3], and Google's Bard[4], and Gemini[5] have made significant gains on the technological market. Among these, ChatGPT has been particularly noteworthy, amassing an impressive 100 million users within just three months of its launch, thereby establishing itself as one of the fastest-growing online platforms to date (Ray 2023). Having been extensively trained on vast datasets, these chatbots utilise machine learning (ML) and natural language processing (NLP). This rigorous training regimen enables them to accurately model complex language patterns, user intentions, and subsequently respond to queries. As a result, these chatbots offer interactions that are not only more precise and refined but also capable of adaptation. They leverage insights from prior interactions to continually enhance their conversational output.

The main issue arises from anthropomorphising these capabilities of LLMs, which just reinforces what is actually an illusion. In order not to fall for the banal deception, we need to remind ourselves that LLMs are fundamentally different from us, as Shanahan (2024) points out. At its core, ChatGPT and LLMs are nothing other than 'bullshit' machines (Hicks, Humphries, and Slater 2024) because they lack self-awareness and knowledge about the world. Even to be able to lie, you must know what you're lying about and need to be able to know the truth-value of your statements (Frankfurt 2005).

Shortly after its release, ChatGPT raised numerous concerns (Borji 2023; Vock 2022; Tripathi 2023), such as providing erroneous information (Borji 2023), exhibiting discriminatory behaviour (Vock 2022), and engaging in inappropriate speech and conduct (Tripathi 2023). The capability of ChatGPT to provide information and interact with users, while impressive, also presents opportunities for misinformation, whether through the limitations of its training data or the inherent biases within these datasets. While prior research has highlighted these issues, the nuanced ways in which users perceive and respond to deceptive outputs remain underexplored.

Addressing this gap is essential for several reasons. First, user trust in AI-driven systems is dynamic—deception can erode trust, but well-calibrated interactions can restore or even enhance it. Understanding the conditions under which users perceive ChatGPT's outputs as deceptive can help refine AI design to promote informed and cautious use. Second, deceptive responses vary across different contexts, making it crucial to investigate where and how misinformation emerges most frequently. Third, the question of responsibility—whether AI developers, hosting platforms, or users themselves bear the burden of mitigating deceptive outputs—has profound implications for AI governance and regulatory policy.

To explore these issues, we conducted a two-part mixed-methods study aimed at investigating how perceptions of banal deception in ChatGPT outputs shape users' behavioural responses and trust, with implications for the design of AI governance frameworks.

[2] https://claude.ai/
[3] https://www.bing.com/
[4] https://bard.google.com/
[5] https://gemini.google.com/app

Study 1 used a survey (n = 220) to identify common types and contexts of perceived deception and to examine how these influence trust, usage, and responsibility attribution (RQ1–RQ3).

RQ1 What are the most common types of perceived deceptive behaviour of ChatGPT, and in which domain (e.g., research, entertainment) do they predominantly occur?

RQ2 How do users perceive ChatGPT's chat-worthiness and responsibility concerning deception, and how do users respond behaviourally to their perceived ChatGPT's deceptive behaviour?

RQ3 To what extent do demographic characteristics (e.g., age, gender) and behavioural factors (e.g., frequency of use) influence users' perceptions of chat-worthiness and responsibility, as well as their behavioural responses?

RQ1-3 Highlights: Our findings indicate that the most frequent types of deceptive behaviour encountered by users were *overly simplified* (53.64%) and *outdated information* (42.27%), with *research* being the most frequent domain for these occurrences. Our analysis shows that the *perceived frequency of deception* impacts users' perceived *chat-worthiness* of ChatGPT without being swayed by personal factors. *Responsibility* for deception is influenced only by *educational level* and *perceived frequency of deception. Behavioural responses*, however, are determined by a mix of demographics (*gender, age*) and other factors (*knowledge, verification tendency, and chat-worthiness*), highlighting a multifaceted set of determinants.

While Study 1 offers valuable insights into the types and contexts of deceptive behaviour perceived by users, as well as their associated behavioural responses (RQ1–RQ3), these findings remain surface-level. Survey data alone could not capture the underlying reasoning, ethical reflections, or trust dynamics that shape how users interpret and respond to deceptive outputs. Prior research suggests that trust in AI systems is context-dependent, emotionally mediated, and influenced by users' mental models and perceived agency of the system (Hancock et al. 2011; Ehsan et al. 2021; Lee and See 2004). Moreover, understanding responsibility attribution in human–AI interactions requires qualitative exploration of users' moral intuitions and sociotechnical expectations (Mittelstadt et al. 2016; Coeckelbergh 2020).

To build on the patterns observed in Study 1 and to examine deeper cognitive and moral dimensions of user perception, we conducted Study 2. This qualitative study aimed to explore how users interpret deceptive outputs, how these interpretations shape long-term trust and behavioural adaptation, and how responsibility and regulatory expectations are assigned (RQ4–RQ5). Using semi-structured interviews with 12 participants drawn from the original survey cohort, Study 2 provides a contextualised understanding of users' lived experiences with deception in ChatGPT and complements the behavioural trends identified in the survey.

RQ4 How do users' experiences with deceptive responses from ChatGPT influence their trust and reliance on the technology, and what methods do they employ to manage these situations?

RQ5 What are users' perspectives on the need for regulatory measures and improvements for ChatGPT, and who do they believe should be held responsible for managing the risks associated with deceptive responses?

RQ4-5 Highlights: Study 2 revealed nuanced insights into users' mental models (Johnson-Laird 2010) and experiences with ChatGPT, emphasising a blend of daily and professional utilisation. Participants reported a generally positive outlook on ChatGPT's conversational capabilities, highlighting its efficiency and utility over traditional tools, yet also expressed concerns over its potential for deception and the ethical implications of its use. Specifically, when encountering deceptive information, there seems to be a notable shift in users' trust levels and attitudes towards ChatGPT. Initially, some participants displayed a low trust level, which either increased upon recognising ChatGPT's advantages or decreased after participants noticed inaccuracies. This led participants to take a more cautious approach when using the technology, which indicates the important role that accuracy, reliability, and explanatory transparency play in shaping user trust.

General Highlights: Our findings from Study 1 and 2 emphasise the complex dynamics of responsibility for ChatGPT's deceptive outputs, with participants attributing responsibility to developers, hosting platforms, and, to

a lesser extent, users of the technology. The results also indicate a consensus on the need for enhanced verification strategies, user education, and regulatory frameworks aimed at mitigating the risks associated with deceptive information. Finally, our study's results highlight the need to address ethical standpoints in the development and use of AI technologies like ChatGPT, advocating for a balanced approach that considers user empowerment, technological improvements, and robust safeguard strategies to enhance trustworthiness and mitigate potential harm.

The paper is structured as follows: In §3 we describe the method and the results from Study 1. In §4, we describe the method and results from Study 2. Then, in §5, we discuss, integrate, and summarise the overall results and insights from both studies. After that, §2 presents the related work in the area of Deceptive AI & Society and contextualises our approach within this area of research. Finally, in §5.5 we discuss future directions in communicative AI agent technologies, and in §6 we conclude the paper. All study materials, including survey questions, interview protocols, and codebooks, are publicly available in the OSF repository [6] for open access.

2 Related Work

In this section, we discuss the related work at the intersection of Deceptive AI & Society (Sarkadi 2023b). Deceptive AI and society research ranges from the more recent studies that capture the relation between LLMs and deception, to the original idea of building a socio-cognitive theory of trust and deception proposed by Castelfranchi and Tan (2001).

The doubt of whether a given technology can be wholly beneficial without any accompanying drawbacks still persists. While explicit failures are easily observable by human eyes, implicit errors are harder to identify and/or fact-check. On one hand, it may be tolerable for chatbots to generate nonsensical responses that merely frustrate users. On the other hand, the possibility of ChatGPT producing *misleading* and *deceptive* information is a matter of serious concern (Nolan 2023; Tiffany and Stuart A. 2023), especially if adopted on a large scale to offer services to users or in safety-critical systems.

Deceptive information through LLM 'hallucination' can have adverse impacts on users who are not equipped to distinguish 'fact' from 'fiction' (Rohrbach et al. 2018; Yijun Xiao and Wang 2021). The dangers that ensue from the use of banal deceptive AI range from LLM being used as tool by other humans to manipulate individuals to causing real harm, and in the extreme, may even result in broader societal ramifications, such as a lack of shared trust among community members and governmental institutions. To better understand and contextualise these risks, prior research has proposed various frameworks to classify different types of deceptive information. For example, Wardle and Derakhshan (2017) distinguishes between misinformation (false information shared without the intent to cause harm), disinformation (false information shared deliberately to mislead), and malinformation (accurate information used with harmful intent). Similarly, Tandoc Jr, Lim, and Ling (2018) reviews definitions and typologies of fake news, identifying forms such as fabricated content, manipulated content, and misleading content. Building on these general frameworks, we developed several common deceptive categories to guide our survey design: "over simplification," "outdated information," "factual inaccuracies," "misleading implications," and "fabricated stories" (see Figure 2). The survey then focused on how users perceive and encounter these predefined forms.

A significant contribution in the area of chatbot technologies is the empirical research conducted by McGuire et al. (2023), who examined user reactions to deceptive behaviours in chatbots. Their findings suggest that users can often fail to recognise deceptive cues, leading to misplaced trust in AI systems. Similarly, Ehsan et al. (2021) focused on the impact of transparency mechanisms in mitigating the effects of trust, indicating that clear communication about an AI's capabilities and limitations can enhance user discernment.

[6] https://osf.io/c7upq/

Pacchiardi et al. (2023) specifically addresses the challenges posed by LLMs, including their ability to generate plausible yet factually incorrect or misleading information. This study emphasises the need for improved detection mechanisms and develops a detector works by asking a predefined set of unrelated follow-up questions.

Going back to balancing our doubts about the threats and benefits of deceptive AI technology in society, we must also emphasise ongoing research that aims to delve deeper into how such technologies can be beneficial and how human-AI interactions work.

In the sub-area of AI called argument mining, a line of work has been to detect deceptive arguments in political debates and contexts using BERT-style systems (Delobelle et al. 2020; Goffredo, Cabrio, et al. 2023). A similar line of work has focused on identifying fallacies and hate speech (Goffredo, Espinoza, et al. 2023).

The study of human perceptions of deceptive AI behaviours has been studied across various of domains such as in linguistics, motor, and social contexts.

In human-robot and human-AI interaction (HRI and HAI), Dragan, Holladay, and Srinivasa (2014) explored how humans perceived the presence of deceptive intentions based on pre-calibrated motions of robotic arms. In the same research area, Chakraborti and Kambhampati (2019) studied how acceptable AI-generated lies were in human-AI teaming search & rescue scenarios. Furthermore, Sarkadi, Mei, and Awad (2023) explored how AI agents are perceived compared to humans when it comes to job roles that involve deception in various agent-agent interactions, including human-AI teaming. Human-AI interaction scenarios also involve the phenomenon where different AI agent strategies can increase human willingness to deceive (Mell, G. M. Lucas, and Gratch 2018). In particular, such effects can be observed in Human-AI negotiation settings (Mell, G. Lucas, et al. 2020; Jahan and Mell 2024).

Two crucial abilities of deceptive AI technologies missing from LLMs are reasoning and planning (Sarkadi 2021; Valmeekam et al. 2023). In the area of AI planning research, an important line of work has looked at extended goal recognition (Masters, Kirley, and Smith 2021) and deceptive path planning (Masters and Sardina 2017; Price et al. 2023). Regarding reasoning, Sarkadi, Panisson, et al. (2019) have explored how AI agents can use abductive and practical reasoning with ToM to cause desired false beliefs in other agents. In the area of multi-agent reinforcement learning (MARL), deceptive AI has been studied in different setups. Piazza and Behzadan (2023) looked at how localised models of Theory-of-Mind can be used to distinguish between cooperative and deceptive AI agent communication.

Another area where Deceptive AI has been studied is in the one of Artificial Societies and Simulation. Specifically related to Deceptive AI & Society, the work of Sarkadi, Rutherford, et al. (2021) have shed light on how deception evolves in human-AI agent societies; on how these societies can self-organise in the face of deception to re-establish cooperative communication through social learning and System 2 type of critical thinking and investigation mechanisms (Sarkadi 2024); on how the presence of deception triggers an arms-race in Theory-of-Mind between deceivers and investigators (Sarkadi 2023a) - a similar result is observed in the MARL approach where ToM is used as a model for Inverse Reinforcement Learning proposed by Alon et al. (2023); and on how competition between agents creates evolutionary pressures to make deception a stable strategy (Sarkadi and Lewis 2024).

Deceptive AI research is only starting to gain traction as a subarea of AI. Several workshops on the topic have been organised in the past years, including the Deceptive and Counter-Deceptive Machines AAAI Fall Symposium[7], the Machine Deception Workshop at NeurIPS[8], the 1st and 2nd International Workshops on Deceptive AI co-located with ECAI 2020 and IJCAI 2021[9], and more recently, the Rebellion and Disobedience of Artificial Agents Workshop Series co-located with AAMAS [10].

[7] https://aaai.org/proceeding/03-fall-2015/
[8] https://www.machinedeception.com/
[9] https://sites.google.com/view/deceptai2021, see proc. in (Sarkadi, Wright, et al. 2021)
[10] https://sites.google.com/view/rad-ai/home

Overall, there's a common thread in all Deceptive AI research, namely that of aiming to better capture what deception is in relation to AI technology, and how it can be used for the good of society rather than increasing risk. As Coeckelbergh (2018) notes, understanding deceptive AI is not just about the technology and engineering of computational systems, but about an overarching narrative about the politics of technology, the power relations and structures that drive technology, and, last but not least, how these play into human cultures and psychological biases. In other words, Deceptive AI needs to be understood as part of an ecosystem (Zhan, Xu, and Sarkadi 2023).

3 Study 1: Online Survey

This study addresses questions RQ1-3, focusing on an in-depth examination of user opinions concerning deceptive behaviours exhibited by ChatGPT. More precisely, Study 1 collects user responses and insights about their experiences with ChatGPT, with a particular emphasis on identifying instances of deceptive behaviour encountered during interactions, such as types of deception and the specific domains where these behaviours commonly occur.

3.1 Method

In this section, we conducted a survey study with 220 participants from both the UK and US. We provide an overview of the data collection and Chi-Square tests, and post-hoc analysis methodology. This study was reviewed and approved by our institution's IRB. To investigate users' perceptions of different forms of deceptive information, we designed the survey with several predefined deception categories. These categories, "over simplification," "outdated information," "factual inaccuracies," "misleading implications," and "fabricated stories" — were developed by the research team based on common patterns of online deceptive practices widely discussed in prior literature (Wardle and Derakhshan 2017; Tandoc Jr, Lim, and Ling 2018). The goal was to capture realistic scenarios that users may encounter online. Participants were asked to indicate whether, and how often, they perceived these forms of deception in their interactions with online content.

3.1.1 ***Participants.*** We recruited participants via Prolific[11]. Using a screening survey, we selected 220 participants[12] who met the following criteria: (a) engagement with ChatGPT in the past six months, (b) experience with deceptive responses during their interactions, and (c) residence in either the UK or the USA. Choosing participants from the US and the UK for the study on ChatGPT's deception responses is justified by their high English proficiency and significant digital literacy, which ensures accurate engagement with AI. These countries' advanced technology adoption and established regulatory frameworks provide a pertinent backdrop for exploring AI interactions and user expectations, offering a comprehensive view of the impact of deceptive AI responses within a Western context. To ensure data quality, we recruited *high-reputation* participants with at least 100 submissions and an approval rate of 95% or more on the Prolific recruitment platform (Peer, Vosgerau, and Acquisti 2014; Such et al. 2017). We obtained valid data from 220 participants. See participant demographics summarised in Table 1, and the demographics were automatically collected by the Prolific platform.

3.1.2 ***Instrumentation & Procedure.*** Our survey was created and hosted on Qualtrics[13]. Initially, participants received an information sheet explaining the purpose of the study, the nature of participation, and confidentiality assurances. This was followed by a consent form, which participants completed to confirm their willingness to participate in the study. The main body of the survey begins with participants' general use of ChatGPT, including the version they use, their frequency of use, and their perceived frequency of receiving deceptive responses. Subsequently, the survey examined the deceptive responses participants believed they had encountered. Given

[11] https://www.prolific.co/

[12] A priori power analysis using G*Power indicated that a minimum sample size of 108 was needed to detect a medium effect with $\alpha = 0.05$, power = 0.80, and df = 2. The final sample size of 220 exceeded this requirement.

[13] https://www.qualtrics.com/

Table 1. Demographics of survey participants.

		#Participants
Gender	Female	108
	Male	110
	Prefer not to answer	2
Age	18-24	41
	25-34	86
	35-44	53
	45-54	27
	55-64	8
	65+	5
Employment Details	Full-time employment	130
	Full-time student	12
	Part-time employment/student	35
	Not employed, job seeking	17
	Not employed, not seeking	14
	Others	12
Education	Middle School	1
	High school	31
	Sixth-form college/school	41
	HND; or University	94
	Postgraduate school	40
	Doctorate	12
	Prefer not to answer	1
Income	Low	107
	Middle	79
	High	27
	Prefer not to answer	7
Total		220

the absence of a well-established taxonomy for deceptive responses, the research team discussed and proposed categories such as outdated information, factual inaccuracies, and misleading information. An open-text field was also provided to allow participants to describe any deceptive responses they had perceived. Finally, participants responded to questions regarding their fact-checking behaviour and their perceptions of ChatGPT, including its perceived chat-worthiness[14] and other relevant aspects. The survey concluded with questions regarding participants' demographics and their willingness to be considered for a follow-up in-depth interview. On average, the survey took about 4 minutes to complete, and we compensated participants at £16 per hour.

Two pilot studies (N=2 each) were conducted to refine our survey instrument. These studies focused on assessing question clarity, layout understanding, and survey logic effectiveness. Feedback from these pilots led to adjustments in question wording and survey design. Data from the pilot studies were used solely for refinement purposes and excluded from the final analysis.

[14]Note that chat-worthiness refers to a user's subjective judgment of whether chatting or engaging in conversation with ChatGPT is worthwhile, based on their overall impression of the value, relevance or usefulness of such interactions.

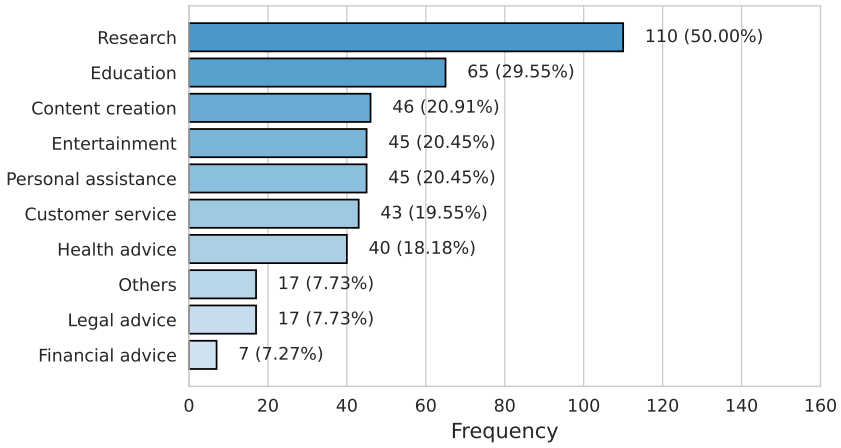

Fig. 1. Common Contexts for Perceived Deceptive Behaviour

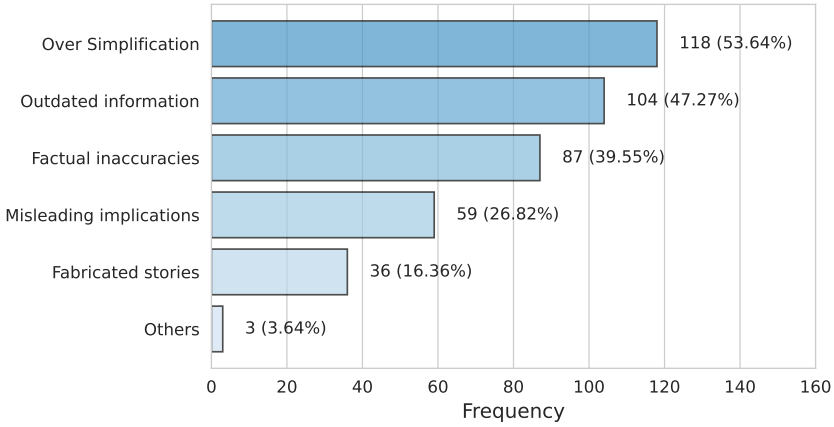

Fig. 2. Common Forms of Perceived Deceptive Behaviour

3.1.3 ***Data Analysis.*** We first conducted descriptive statistics to address RQ1. This was followed by employing a Chi-square test (McHugh 2013) to investigate whether and how demographics and personal factors - encompassing users' knowledge of LLMs, frequency of ChatGPT usage, frequency of receiving deceptive responses, and the frequency of verifying ChatGPT's responses - influence perceptions of ChatGPT's chat-worthiness, responsibility, and users' post-behaviour. Given the multiple comparisons involved, we applied a Bonferroni correction to control for experimenter-wise error rates, adjusting the significance thresholds ($p < 0.005$) accordingly.

3.2 Study 1 Results

3.2.1 ***Common Forms and Contexts of Deception.*** We illustrate the common deceptive categories in Figure 2, including "over simplification", outdated information", factual inaccuracies", misleading implications", and fabricated stories". We observed that the most frequent form of perceived deceptive response encountered by ChatGPT users is *Over simplifications*, reported by 53.64% of participants. This category is **the only one**

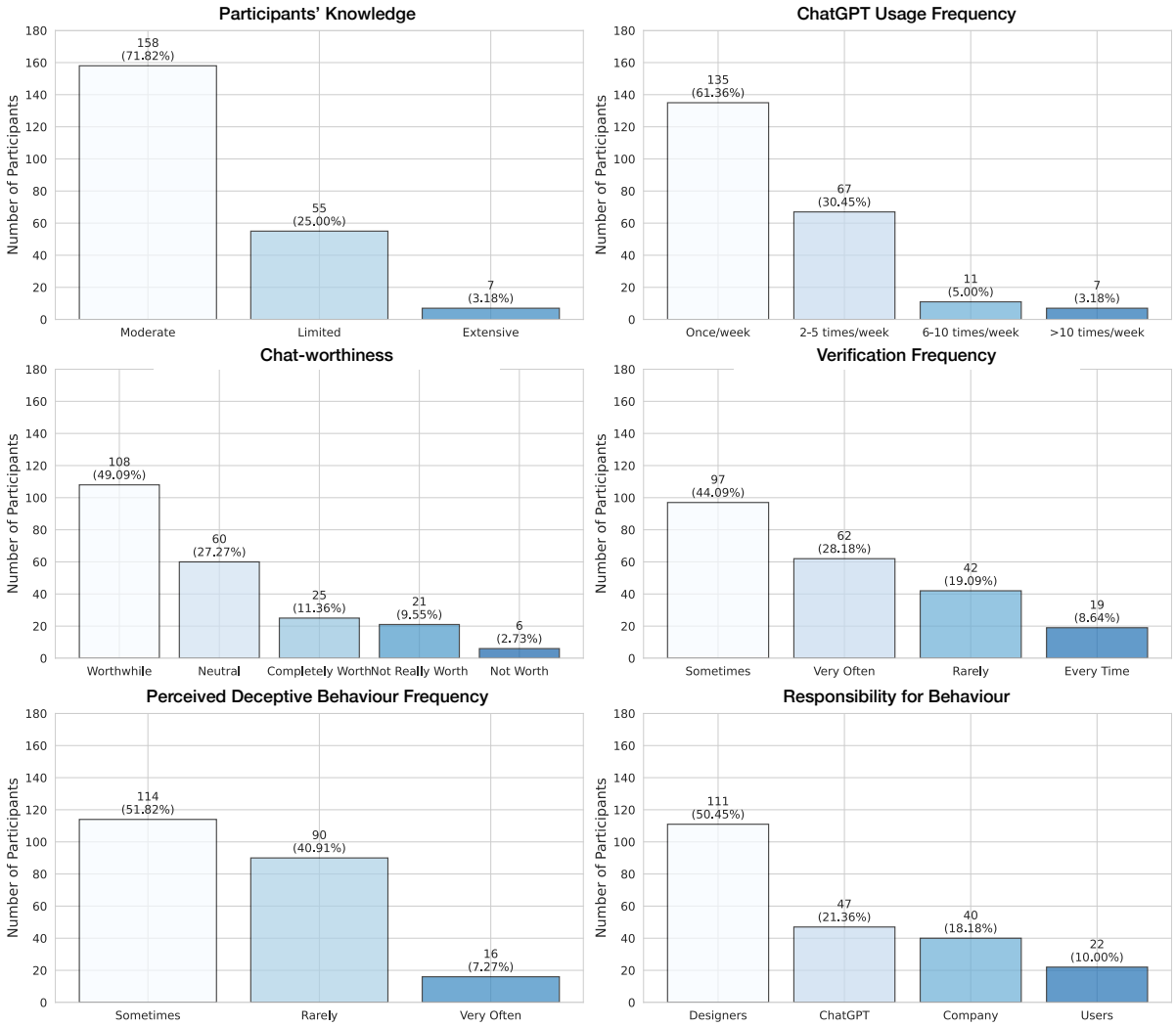

Fig. 3. Descriptive analysis of survey responses regarding: (1) Participants' Knowledge, (2) ChatGPT Usage Frequency, (3) Chat-worthiness, (4) Verification Frequency, (5) Perceived Deceptive Behaviour Frequency, and (6) Responsibility for Behaviour.

surpassing the 50% threshold. This implies that a substantial amount of the information may not provide a thorough comprehension of the subject topic, potentially leading to misunderstandings or misinterpretations.

Figure 1 shows the common areas where participants perceive that they encounter deceptive behaviours in ChatGPT, including education, entertainment, personal assistance, customer service, health, financial advice, and other legal advice. It indicates that a significant proportion of participants, about half, experienced deceptive responses while discussing *research-related* topics with ChatGPT. Furthermore, Education ranks as the second most common context (29.55%) for receiving perceived deceptive information.

3.2.2 ***Descriptive Analysis of Personal Factors.*** Regarding their knowledge of AI, 71.82% of participants rated their knowledge as *moderate*, and 3.18% considered themselves experts. Moreover, the majority of users do not frequently use ChatGPT, with over half (61.36%) utilising it just *once a week*. The second-largest group of participants (30.45%) uses it between *2-5 times a week*. Only 5% of users engage with it *6-10 times* weekly, while a mere 3.18% do so *more than 10 times a week*. When it comes to the frequency of using ChatGPT, a majority of participants, constituting 51.82%, reported that they *sometimes* receive deceptive information, whereas 40.91% of participants indicated that they *rarely* encountered deceptive information. A minimal fraction of the participants, approximately 7.27%, stated that deceptive behaviours appeared *very frequently*. In terms of verifying ChatGPT's responses, 44.09% reported doing so *occasionally*, while 8.64% *always* checked the accuracy of the information provided.

3.2.3 ***Users' Opinions on ChatGPT's Chat-worthiness, Responsibility for Deception, and Their Behaviour Change.*** In evaluating the chat-worthiness of ChatGPT, we utilised a scale from 1 to 5, where 1 signifies the lowest value, indicating it is not worth chatting with at all, and 5 indicates the highest value, suggesting it is entirely worthwhile. The majority of respondents, over 60.45%, rated their experience *positively*. Only a small fraction, 2.73%, felt it was not worth chatting with, giving it the lowest rating of 1.

We were also keen to investigate whether users would continue to use ChatGPT after receiving incorrect answers. Surprisingly, we found that after experiencing deceptive behaviour, 54.1% of participants continued to use ChatGPT, 38.6% chose to reduce their usage, and only 7.3% decided to stop using it altogether.

In terms of the question related to responsibility for ChatGPT's potentially misleading outputs, a majority of 50.45% pointed to its *designers or creators*. Approximately 21.36% attributed the responsibility to *ChatGPT itself*, while only 10% believed that users should be accountable.

Table 2. Results of the Chi-square test (where '-' indicates non-applicability and empty cells represent non-significant outcomes, hence not included in this table).

Factors	**Chat-worthiness**			**Responsibility**			**Behavioral Response**		
	χ^2	p	Sig.?	χ^2	p	Sig.?	χ^2	p	Sig.?
Gender							10.090	0.006	Y
Age							25.140	0.005	Y
Employment									
Education									
Income									
Knowledge							14.636	0.006	Y
Deception Fre.	22.084	0.005	Y	19.355	0.004	Y			
Use Fre.	34.208	<0.001	Y				-	-	-
Verification	-	-	-				20.826	0.002	Y
Chat-worthiness	-	-	-	-	-	-	98.514	<0.001	Y

3.2.4 ***The impact of participants' demographics and personal factors on Chat-Worthiness, Responsibility, and Behavioural Response.*** As demonstrated in Table 2, the differences of *Behavioural Response* among different *genders* are evident in the comparative percentages. Males predominantly chose to keep using ChatGPT, while a significantly higher percentage of females (47.92%) opted to reduce its usage. As for various age groups, those aged 18-24 demonstrate a significantly higher likelihood of continuing to use ChatGPT in comparison to other age brackets.

Regarding the impacts caused by personal factors, the *participants' perceived frequency of encountering deceptive responses* significantly affects participants' perceived *chat-worthiness* of ChatGPT and their views on *responsibility*. Notably, individuals who receive deceptive responses *very often* tend to view ChatGPT as *slightly unworthy* of chatting and are more likely to attribute responsibility for these deceptions to the *designers*. In contrast, those who *rarely* encounter such responses tend to consider it *slightly worth* chatting with and are inclined to hold the *company* accountable. These findings align with general expectations and common understanding. The *frequency of using* ChatGPT is significantly correlated with users' perception of its *worthiness*. However, the most frequent participants (over 10 times a week) are inclined to view ChatGPT as "slightly not worth talking to". Conversely, those using it 6-10 times a week lean towards seeing it as "slightly worth talking to". The least frequent users (once a week) predominantly opt for a *neutral* stance in their assessment. For factors that significantly affect participants post-behaviour, individuals possessing *extensive and expert knowledge* of LLMs show a significant preference for *keeping* their use of ChatGPT. In contrast, those with *lower* level knowledge are notably more inclined to *reduce* their usage. Individuals who *sometimes* verify responses generated by ChatGPT *keep* using the service, while those that verify their responses *every time* are inclined to *reduce* their usage.

4 Study 2: Semi-structured Interview Study

While Study 1 focuses on answering research questions RQ1-3 in a descriptive manner, we aim to explore further how users perceive their encounters with ChatGPT's deceptive behaviours and how these experiences impact their usage, trust, and future preferences (RQ4 and RQ5). Study 2 provided insights into users' mental models and experiences with ChatGPT, showing its use for both daily and professional purposes. To explore users' perceptions of different forms of deceptive information generated by ChatGPT, in Study 1, we designed a survey using predefined deception categories. To validate this predefined framework and allow for new insights, Study 2 we aim to provide an open-ended opportunity for participants to describe their experiences in their own words, helping us assess whether the predefined categories were sufficient and whether any new types emerged.

4.1 Method

In this section, we conducted semi-structured interviews with 12 participants from both the UK and the US. We provide an overview of the data collection using thematic analysis. This study was reviewed and approved by our institution's IRB.

4.1.1 ***Participants.*** From the original pool of 220 survey respondents, 14 participants (12 for final analysis and 2 for pilots) were chosen for the in-depth interviews. This approach facilitated seamless progression to more detailed explorations during the interviews, leveraging the participants' pre-established familiarity with the survey themes. Recruiting interviewees from our survey respondents not only streamlined the research process by eliminating the need for a new recruitment phase but also minimized potential sample bias. Selection was based on their willingness to participate further, and a stratified sampling method was employed to balance the participants' variances, especially in significant factors shown in Table 2. For instance, we found that participants' *age* and *gender* significantly influence their perceptions of post-behaviour. To comprehensively investigate this phenomenon, we recruited participants who exhibited a wide range of age groups and genders in our survey study. The demographics of these participants are summarised in Table 3.

Potential participants were approached through the Prolific platform. All interviews were conducted virtually via Zoom[15], with the main interactions averaging approximately 30 minutes in duration. This excludes the time spent on introductions and explaining the purpose of the study. With prior consent from the participants, all sessions were audio-recorded. Each participant was compensated at £20 per hour through Prolific. The interviews

[15] https://zoom.us/

Table 3. Demographics of interview participants. Note '-' indicates that this participant had retired before the release of ChatGPT and therefore was not eligible to use ChatGPT at professional work.

PID	Age	Gender	Knowledge	Job	Usage	
					Professional work	Daily life
P1	18-24	Male	Extensive	NLP researcher	✓	✓
P2	25-34	Female	Extensive	University lecturer in Information	✓	✓
P3	45-54	Female	Limited	Chef	×	✓
P4	65+	Male	Moderate	Retired	-	✓
P5	35-44	Male	Moderate	Solicitor	✓	✓
P6	25-34	Male	Moderate	Clinical researcher	✓	✓
P7	25-34	Male	Extensive	Robotics Company Engineer	✓	✓
P8	25-34	Female	Moderate	Marketing manager	✓	✓
P9	45-54	Male	Limited	University professor in Music theory	✓	✓
P10	35-44	Male	Limited	Civil servant	✓	✓
P11	18-24	Female	Moderate	Software tester	×	✓
P12	25-34	Female	Moderate	Student	✓	✓

were conducted from January to February 2024. Among the 12 participants interviewed, there were 7 (58.3%) males and 5 (41.7%) females, with their ages spanning almost all age groups: 2 (16.7%) individuals aged 18-24, 5 (41.7%) individuals aged 25-34, 2 (16.7%) individuals aged 35-44, 2 (16.7%) individuals aged 45-54, and 1 (8.3%) individual aged 65 or above. Among these participants, 3 (25%) possess extensive knowledge of ChatGPT, 6 (50%) have moderate knowledge, and 3 (25%) have only limited knowledge.

4.1.2 ***Interview Protocol.*** The interview protocol was semi-structured, known for enabling detailed and comparable qualitative data (Knott et al. 2022). During each interview, researchers strategically posed opportunistic follow-up questions as necessary, aiming to comprehensively capture the participants' experiences. The interview script contains questions on the following topics:

(1) Participants' usage of ChatGPT in both personal and professional settings, their integration of its assistance with their own skills, and their motivations for using it. It examines specific tasks where ChatGPT is crucial, their knowledge of its capabilities, and comparisons with other tools like search engines.
(2) Participants' experience with deceptive or inaccurate responses from ChatGPT, their methods of handling such situations, and the impact on their perceptions of ChatGPT's reliability and subsequent behaviour.
(3) Factors affecting trust in ChatGPT and participants' views on its future reliability. This includes responsibility for deceptive responses and potential risks of over-reliance, especially for vulnerable groups.
(4) Participants' preferences and expectations for regulatory measures and improvements to ChatGPT.

To refine the protocol, we initiated the process with a pilot study involving two participants following the same selection procedure. This enabled us to refine the interview structure, ensuring each question was clear, understandable, and effectively designed to elicit the targeted information. The pilot interviews were excluded from the final dataset for analysis.

4.1.3 ***Data Analysis.*** We performed an inductive thematic analysis (Braun and Clarke 2006) to process participant responses. We started with two researchers independently coding the first interview transcript to identify salient codes, thereby establishing an initial codebook. This preliminary codebook was then collaboratively refined during the analysis of the second interview transcript, where the researchers engaged in a detailed discussion to reach a consensus on the codes that were applied to the remainder of the interview transcripts. Then, the two researchers independently coded subsequent transcripts, when new codes emerged, the researchers met to discuss these new findings and, where necessary, made amendments to the codebook. This iterative

process continued until no new codes were identified, indicating a point of code saturation, which occurred after analysing seven transcripts. Upon completing the coding of all interviews, the researchers collectively reviewed and deliberated on the potential themes. This collaborative review process was instrumental in ensuring the thematic saturation and in achieving a consensus on the final themes. To ensure a high level of inter-coder reliability throughout the study, Cohen's kappa statistics (Fleiss, Levin, and Paik 2013) were computed for each interview transcript, and the final average is .84, indicating *substantial agreement* between the researchers. This suggests that the coding scheme was applied consistently, lending credibility to the thematic analysis conducted.

4.2 Study 2 Results

The qualitative findings are reported below. We have edited the reported quotes to remove filler words e.g., *"umm", "like", "ah-ha"*) with *Hemingway Editor*[16] used to indicate where quotes have been condensed for brevity.

4.2.1 ***Mental Models of & Experiences with ChatGPT.*** This section reports users' interactions with ChatGPT, focusing on their uses, motivations, and general attitudes.

Daily vs. Professional Purposes. During interviews, participants described using ChatGPT for both personal and professional purposes. All 12 integrated it into daily life, with 9 (75%) also employing it professionally. Daily uses included artwork creation, writing assistance, idea generation, and social interaction, while professional uses spanned academic content creation, newsletter generation, automated grading, medical QA, legal advice, coding, and information integration in fields like education, healthcare, and law. Professional users often leveraged specialised adaptations of ChatGPT. For instance, P6 (Clinical Researcher) used a customised institutional version, while P1 (NLP Researcher) explored multiple LLMs beyond standard GPT models. Most participants, except P1, were motivated by curiosity about ChatGPT's capabilities or social influence ('herd mentality') rather than a specific professional necessity.

Positive Perspectives. Participants highlighted ChatGPT's effectiveness, efficiency, and superiority over traditional tools like search engines and translation software. They noted its growing adoption in academia and industry, with P2 mentioning universities' interest in generative AI and P6 describing their company's custom ChatGPT, *"Our company has also created its own version of ChatGPT. We have a dedicated team responsible for its development and maintenance. Employees are taught and encouraged to use it for work-related tasks, like finding information about a specific drug, [...]"*. Many participants also believed ChatGPT would improve over time, reducing deception issues.

Negative Perspectives. Compared to positive attitudes, there is a predominant emphasis on negative perspectives expressed by participants. This includes 1) inherent negative perceptions regarding AI technology, 2) negative consequences of using ChatGPT that have already occurred and concerns about potential future ones. Note that we separate out the negative consequences that ChatGPT's deceptive responses cause or may cause and discuss them in §4.2.2, and 3) general feelings that current ChatGPT is *"limited in its capabilities"* [P6, P7, P12] or *"not ready yet"* [P2]. Among them, we observed that there is a concern among participants regarding the privacy and safety implications of using ChatGPT in general [P1, P2, P6, P12]. Participants questioned the security of personal data and the reliability of ChatGPT's aggregated information.

Knowledge of ChatGPT.. Participants exhibited varying levels of understanding and assumptions about ChatGPT's operations, ranging from its source of information to its capabilities and limitations. Interestingly, despite the varying levels of self-reported knowledge among users (see Table3), their understanding of ChatGPT's operations, as discussed during the interviews, was remarkably consistent. Several participants [P6, P7, P8, P9] described ChatGPT as a data mining or scraping tool, leveraging large repositories of internet data, including a

[16] https://hemingwayapp.com/

mix of pre-existing knowledge and generative capabilities based on predictive algorithms. There was a recognition of the vast amount of data ChatGPT has access to, including digitised books and potentially internet forums, although there was uncertainty about its access to subscription-based journals and books. Only one participant [P3] believed that ChatGPT's information source was based solely on user input.

4.2.2 ***Deception and User Reactions.*** In this section, we summarise user experiences with deceptive information from ChatGPT, focusing on their reactions and perceptions of these encounters.

Table 4. Participant Quotes Illustrating Types of Perceived Deception

Perceived Deception Type	**Description**	**Example Quotes**
Oversimplified answers	Provided oversimplified or incomplete answers.	*(P2) "It wouldn't give me the actual content that would go into the lecture.";* *(P9) "Incomplete answers I have to double check.";* *(P10) "It just came back with yes or no or very little information."*
Partially incorrect information	Responses are partially correct and partially wrong.	*(P6) "It messed the reference up.";* *(P8) "Some names were real, some made up.";* *(P9) "Portions correct, portions wildly incorrect.";* *(P11) "One paper didn't exist, one had wrong year."*
Misleading information	Misleading or logically inconsistent answers.	*(P2) "A lot of the answers were misleading.";* *(P5) "GPT suggests things that don't make sense.";* *(P5) "Crosses legal backgrounds from different jurisdictions."*
Factual inaccuracies	Incorrect logic, maths, or coding outputs.	*(P2) "It wasn't good with logic or maths.";* *(P4) "Never got code to run.";* *(P10) "Football team's trophies wrong.";* *(P12) "Sometimes it throws errors and fails."*
Outdated information	Outdated responses due to limited training cut-off.	*(P2) "It can only give information up to 2021.";* *(P4) "Anything recent, clueless.";* *(P8) "Sometimes makes stuff up about current events."*
Fabricated information	Makes up content or names.	*(P4) "Made up a poem that didn't exist.";* *(P6) "Went down a weird, made-up route.";* *(P8) "Listed made-up names that weren't real thought leaders."*
Unable to imitate writing style	Fails to match user writing style.	*(P4) "Never quite matched the newsletter style consistently."*

Deceptive Information Received by Participants. Most of the deceptive information described by participants aligns with the categories presented in Figure 2. To illustrate this alignment, representative quotes from the interviews are summarised in Table 4. These examples show how participants' own words confirm the predefined types and also highlight additional nuances, such as ChatGPT's inability to consistently imitate writing styles [P4]. Interestingly, misleading responses generated by ChatGPT are able to manifest themselves both explicitly and implicitly to our participants. For instance, claims made by ChatGPT like *"Vaccines often cause more harm than good"* and *"You can always trust news shared on social media"* are obviously dubious. Whereas what participants

find particularly troublesome is that ChatGPT sometimes *"produces partially correct information or advice, making it extremely difficult to discern truth from falsehood"* [P1, P6-12]. A very interesting example experienced by P8:

> *[...] asked ChatGPT to generate a list of current thought leaders in the marketing industry. And it listed some names and I don't remember who they are now, but some names were real, and some were just like some totally made-up person who wasn't a real thought leader in the marketing industry.*

Response Checking and Behaviour Changing. With regard to the deceptive information provided by ChatGPT, essentially all participants were aware of this problem, with only P5 mentioning that they would perform a detailed check on almost every response given by ChatGPT. P2 indicated that the proactive verification behaviour only began after the first time noticed deceptive information in ChatGPT's response, wherein P2 realised *"indeed ChatGPT can also make mistakes, okay, let me check these solutions to make sure it's reliable."* Most commonly, participants [P1, P2, P4, P6, P8-12] (9/11) decided whether to conduct detailed fact-checking based on the importance of the context in which they want ChatGPT to function. For example, P10 mentioned *"it depends on what you're using it for, I was using it for something like a medical diagnosis like something critical, I think that we want to be checking things, [...]"*. Moreover, a significant obstacle arises when participants pose questions to ChatGPT that fall outside their knowledge or expertise, making it difficult for them to verify responses and identify potential inaccuracies [P5, P7]. P5 even described feeling *"incompetent"* when addressing unfamiliar responses and candidly admitted, *"I would take [ChatGPT's response] as the truth."* This predicament is particularly pronounced among individuals with an inherent inclination to trust, who may lack the motivation to scrutinise the responses further, thus unwittingly accepting inaccurate information.

After experiencing deceptive responses from ChatGPT, participants' subsequent behaviours diverge. 8 participants [P2-4, P7-9, P11, P12] reduced their usage of ChatGPT for the given task. As for the reason, it includes *"I've kind of gone back to Google after an initial enthusiasm"* [P4] and realises this doesn't help and *"might cause messy for serious tasks"* than just for entertainment [P2, P8]. However, the remaining 3 participants overlooked the deceptive responses and maintained their original frequency of use. They justified this decision by explaining that they relied on ChatGPT for very simple tasks, where deceptions were readily identifiable.

Correct Actions and Efficacy of ChatGPT Responses. Participants employed diverse strategies to rectify ChatGPT's responses. Some opted for a straightforward correction by responding with a brief assertion such as *"you are wrong"* without furnishing additional instructions [P3, P5, P8]. Conversely, others also offered guidance or specific requirements in their prompts [P1-4, P7, P8, P10-12], for instance, P4 mentioned *"[...], I was feeding it previous examples of articles I'd written, [...]"* and *"[...] say, this code didn't run, you know, I got this error message, and I put it back to [...]"*. Subsequently, it was observed that a greater number of participants (n = 5) did not experience an improvement in ChatGPT's responses, whereas only 4 participants reported an improvement. It is noteworthy that these improvements were all predicated on participants providing explicit instructions. Participants also specifically mentioned the limitations of ChatGPT in functionalities (modalities) extending beyond text, such as generating images [P1, P3, P4, P12]. When participants provided additional guidance to ChatGPT, they observed that the output *"got incrementally worse"*. To elucidate further, ChatGPT even induced a feeling among users, described as *"It seemed to get confused with further instructions, [...], every time I added something, it just got more murky and, it lost the integrity of what it was trying to be, [...]"*[P3]. P10 stands out as the only individual who does not attempt to correct ChatGPT upon discovering deceptive information, demonstrating indifference towards its accuracy with a remark, *"well, that's wrong. That's kind of the end of it."*

Perceptions on Why ChatGPT Generates Deceptive Responses. But when it comes to the reasons behind the deceptive information, only P1, P5, P7, P8, P11, and P12 offered their speculations. The other participants expressed that they found it strange but were unsure of the specific causes. P5, P8, and P11 noted that their impression of ChatGPT is that its objective appears to be to strive or attempt to provide responses that seem logically

coherent or to assemble elements that sound correct, regardless of the question or request, even though the actual information provided may be inaccurate. P7 posited that ChatGPT may occasionally misinterpret specific words or sequences of words within user prompts. Finally, P12 mentions that maybe the information in the ChatGPT knowledge base or training data is inherently wrong.

Negative Consequences Regarding Deceptive Information. Fortunately, to date, none of the 12 participants have faced any risks or encountered serious consequences as a result of receiving deceptive information. However, most have reported awareness of individuals within their networks who have experienced such issues, or they have expressed concerns regarding the potential harm deceptive information may cause.

Participants [P1, P5, P6, P8-11] expressed concerns about the consequences of specific groups receiving deceptive information from ChatGPT.

To provide more details, we identify the following demographics that participants have deemed particularly susceptible to the deceptive information disseminated by ChatGPT.

- Kids and the Elderly [P5, P6, P9]. In participants' minds, kids and the elderly are particularly vulnerable due to a combination of developmental, cognitive and technological factors. Kids are still developing critical thinking skills and inherently more trusting, while the elderly, might be affected because of the potential sensory and cognitive declines. For instance, P6 thought kids *"do generally trust people. And they're not as sceptical as adults, [...]."*
- (Young) students [P6, P8-11]. Young students are particularly vulnerable to deceptive information from AI technologies like ChatGPT due to their specific need for quick answers and their familiarity and comfort with using such technologies. This demographic's tendency to rely heavily on AI for academic assistance, without adequately verifying the information's accuracy or engaging deeply with the material, heightens their risk of being misled. P9 specially mentioned, that young students' dependence on ChatGPT for quick solutions can atrophy their ability to independently evaluate arguments, and P9 articulated apprehensions regarding the long-term effects of such dependence, positing that *"I suspect the more that we rely on ChatGPT, it's possible that our own skills especially the critical thinking ability will diminish, [...]"*
- Individuals Unable to Afford Healthcare Services [P8]. The participant believes that due to high healthcare costs in the US, people who rarely visit doctors or cannot afford them might use ChatGPT for initial health advice. They contrast this with the UK[17], suggesting the issue might be less severe there, and indicates this behaviour is a result of financial barriers to accessing medical care.
- Non-tech-savvy [P1, 10, P11]. People with no technical background are likely to overtrust AI due to unfamiliarity with technology's limitations and a belief in its infallibility, as highlighted by P1's observation. They might assume the computer's infallibility, believing P10, "*the computer must know; it knows all this other stuff, so it must be right.*" This naivety can lead to uncritical acceptance of potentially inaccurate information as noted by P1 and P11. Additionally, P11 mentioned that individuals with mental health issues are particularly vulnerable, as they may find it even more challenging to discern the reliability of information provided by AI.

4.2.3 ***Responsibility and Trust.*** By asking specific questions, we delve into the complex dynamics of perceived responsibility and trustworthiness concerning ChatGPT's deceptive information. In terms of the perceived responsibilities of various stakeholders, we distinguish between 'developers', defined as employees responsible for technical product development, and the 'hosting platform', the entity or company managing and operating the product. In the context of ChatGPT, 'developers' refers to the engineers, while OpenAI serves as the 'hosting platform' overseeing its deployment and API management.

[17]Hospital treatment is free to people who are "ordinarily resident" in the UK. https://www.jpaget.nhs.uk/patients-visitors/overseas-visitors/information-for-people-seeking-free-nhs-hospital-treatment/

Unintentional Deception and No One Should be Responsible. When questioning responsibility for ChatGPT's deceptive behaviour, several participants [P1, 3-5, P6, P7, P9, P11, P12] (9/12) emphasised that ChatGPT, lacking 'consciousness', does not engage in deception intentionally. And, as such, should not be held responsible for such actions. While acknowledging the *'the 'existence of AI systems intentionally trained by humans to disseminate misleading information, which could be deemed deceptive,"* P5 still believed ChatGPT should not be categorised like that. Furthermore, as described by P7, *"[. . .] It hasn't lied to me. I take it as incompetence. I take it as a lack of knowledge, [...]"*, highlighting it as a limitation of the AI technology behind ChatGPT.

Developers' Responsibility. [P1, P2, P5, P7, P9, P10-12] (8/12) participants attributed the responsibility for ChatGPT's potentially deceptive outputs to its developers, emphasising the developers' crucial role in designing algorithms that are discerning in data verification and source selection from the outset. As creators of ChatGPT, developers are tasked with safeguarding the integrity and accuracy of the content provided, and they are expected to uphold an ethical obligation to ensure ChatGPT's utility and establish its credibility among users.

Hosting Platform's Responsibility. A subset of participants (5/12) identified hosting platforms as responsible parties. They highlighted that these platforms are obligated to guarantee the ethical, legal, and secure deployment of these systems, emphasising the platforms' financial interest in their business models. With P6 mentioned the hosting platforms *"gonna be the people to decide how it operates and what it says"* and in contrast developers *"don't have any control over it"*. This expectation is particularly pronounced when participants compare ChatGPT to widely used consumer products, with P3 stating, *"[...] like if I choose to use eBay or Amazon, I expect them (the platform) to bear responsibility for the content they publish."*

ChatGPT Itself. Only P10 holds the view that the responsibility lies solely with ChatGPT. This perspective stems from their understanding that their engagement is *"directly with ChatGPT"*, from which they receive information. Throughout this interaction, P10 does not take into account the potential involvement of any other entities.

User Themselves. Responsibility attribution to users themselves was less common , with only 4 [P2, P3, P6, P7] participants acknowledging it. They cited a "*lack of sufficient knowledge*" about ChatGPT as a key factor limiting their effective engagement with the system. P3 reflected on this perspective, influenced by their awareness of friends using ChatGPT for advanced tasks, and concluded that their own limited understanding of generative AI technologies prevented them from fully exploiting ChatGPT's potential, stating, "*Nothing wrong with the machine. It's my lack of knowledge that stops me from getting the full experience.*"

Trust Changes after Receiving Deceptive Information. When it comes to participants' perceived trustworthiness of the ChatGPT, especially when our researcher asked if their trust level had changed from the beginning participants started to use ChatGPT and after receiving the deceptive information. We first observed participants trust more along with their use of ChatGPT. To be more specific, initially, some participants [P2, P7] reported a low level of trust, influenced by negative perceptions or lack of understanding of ChatGPT. However, the trust increased later after recognising its advantages [P2, P6, P7], and adjusting expectations accordingly [P10]. Conversely, others maintain a consistent level of scepticism or trust [P1, P3, P9]. For some, trust does not significantly change because they start with realistic expectations about the AI's accuracy and usefulness, particularly in their specific areas of interest or professional needs [P9, P10]. For example, P10 stated *"I wasn't expecting it to be a hundred per cent right [...]"*. Four participants [P4, P8, P11, P12] experienced a decrease in trust when faced with inaccuracies or limitations in the responses. These experiences lead to a more cautious approach to using the technology, including verifying information independently [P11] and adjusting how they use the AI based on its limitations [P8, P12] (*"[...], as trust goes down, I try to steer clear of using it for tasks where it's not doing great." [P12]*).

Factors Affecting Trust. Trust in ChatGPT, as reflected through participant feedback, is shaped by a complex interplay of factors that underscore the nuanced perceptions of its reliability and utility.

- Accuracy and Reliability [P2, P5, P9, P11]. Participants emphasised the critical importance of accurate and reliable information, with inaccuracies significantly undermining trust.
- Explanatory Transparency [P2, P7]. 2 participants highlight the value of clear explanations regarding ChatGPT's reasoning processes as a means to foster trust.
- Content Guidelines and Disclaimers [P5]: P5 pointed out that the implementation of content restrictions and disclaimers, especially on sensitive topics, informs users of ChatGPT's limitations, thus guiding trust.
- Domain-Specific Trust Variations [P1, P2, P4-6, P8]: Trust varies significantly across domains, with participants expressing higher trust in specific areas (e.g., linguistic tasks) and caution in others (e.g., political content).
- Influence of External Perceptions [P10, P11, P12]: Participants expressed that narratives conveyed through media channels, as well as feedback from immediate social circles (including family and friends), have an impact on their level of trust towards ChatGPT.
- Clarity of Source Data and Integration with Tools [P7, P12]: Clear data sources and seamless integration with other tools are key to building trust. P12 noted, "If more tools are integrated, I'd likely use and trust it more."

4.2.4 ***Perceived Future Expectations of ChatGPT.*** Participants shared suggestions and expectations for the future development of ChatGPT to get rid of deceptive behaviours from multiple perspectives.

General & Technical Improvement Needs. Participants underscored the necessity for advancements in ChatGPT to address and reduce incidences of deceptive behaviours and misinformation by enhancing transparency and accuracy, and introducing robust validation processes. Furthermore, several participants expected ChatGPT to evolve beyond its current capabilities as a generative AI model. P8 acknowledged ChatGPT's potential to serve as a *"real-time assistant"*, while P9 and 11 envisage it embodying more human-like qualities, facilitating its application across diverse life aspects, including education and healthcare sectors. P8 elaborated on these expectations by stating, *"ChatGPT is capable of generating responses. However, what it lacks is the ability to act upon these responses. So I would like to see it not only generate responses but also execute actions based on them. That's where we're going to get to."*.

Verification & Safeguard Strategies. Participants in the discussion on verification and safeguard strategies for ChatGPT express a range of views emphasising the shared responsibility between users and developers for verifying information accuracy. Foremost, it is argued that an enhancement of ChatGPT's capabilities to assess its certainty regarding the provided information and to authenticate the origins of its data prior to responding to users would be beneficial and should be prioritised. Many also advocated for the user's responsibility to perform due diligence [P8-11], while also recognising the complexity of placing the entire burden of fact verification on developers. P11 specifically noted, *"you [users]'re the one that agreed to their terms of service to use the platform so you always need to double check any information and should you take any information at face value."* This emphasises the critical responsibility users bear in authenticating the information provided by the platform perceived by the participants. Furthermore, a notable preference for third-party verification emerges [P2, P3, P5, P6, P11, P12], suggesting it could offer unbiased, accurate checks and enhance user trust. However, participants also raised a subsequent concern regarding the credibility of the third-party verification tools themselves and the methods through which their trustworthiness can be substantiated (*"So because we don't fully trust ChatGPT, we're thinking of using a third-party tool for help. But then, do we need to double-check if that third-party tool is even reliable? It's like a never-ending loop."* [P1]).

Empowering Users & Increasing their Self-esteem. This includes educating users, improving user self-protection, and continuous engagement in ChatGPT development and on in-using phase. Central to the dialogue is the imperative for enhanced educational initiatives and user guidance, to foster a deeper understanding of ChatGPT

and the like: *"I think everyone should have some kind of basic guidance or training on using ChatGPT before they use it [...] because it's important to know the type of technology you're using before you deploy it in the field." [P11]*. These considerations are crucial, as highlighted in §4.2.2, for individuals lacking technological proficiency, including children and the elderly. Owing to their limited comprehension of ChatGPT's operational mechanisms, these groups are more vulnerable to adverse outcomes stemming from deceptive information.

The consensus among participants was that users need to be more aware of self-protection when using ChatGPT. Participants expressed the view that users should not place too much trust in ChatGPT, which still has limitations in understanding the nuances of individual situations and can be misleading. Therefore, they expressed concern and disapproval of relying on ChatGPT to make important life decisions: *"I can't just go right on ChatGPT, can I sell my house for example or can I move to a different country? Those are decisions that need to be made by humans, not generative AI."* [P2]. In terms of the dependence on the use of ChatGPT, participants advocated a call for a cautious approach using ChatGPT, as they have concerns that it will destroy people's ability in critical thinking, and accordingly, lose their capabilities in specific tasks or skills. For instance, P12 mentioned that relying on ChatGPT will cause a more severe dependence on the ChatGPT and securely, *"Now when I write, I just let it go and I get lazy myself and don't deliberate on which word or phrase to use as I used to."* P8 also mentioned the importance of critical thinking and the ability to discern the credibility of information in the digital age, particularly in the context of social media and potential misinformation: *"And if people start just trusting you at face value, because they've been told, oh, it's, it's vetted and it's regulated, then they kind of lose that ability to think critically for themselves. And it comes down to even on social media, if someone's posting an article with a misleading headline, like that skill of being able to verify and, you know, have your own critical thinking, it carries over into other aspects of life."*.

Participants pointed out that the current development and evolution of ChatGPT have not fully and seriously considered users' needs and experiences. Therefore, they advocated for a more inclusive, user-focused approach to the improvement of ChatGPT and other AI products. To be more specific, participants believe that it is crucial for companies to conduct user acceptance testing to ensure products meet the required standards for functionality and reliability (*"if they don't get enough people saying, this is really good and it's producing accurate results, then they shouldn't release it" [P4]*). Once the product is in the market, collecting ongoing user feedback and learning from complaints and comments are vital for iterative improvements. P7 suggests that tools for feedback, such as *"rating systems within the product"*, should have a tangible impact on development and enhancements. Furthermore, the discussion identified user studies as a crucial strategy to address significant gaps, including a limited understanding of user requirements and a neglect of ethical considerations. These studies are considered effective in gaining a holistic comprehension of end-user needs, requirements, and feedback. P1 critically noted: "*Without conducting user research, developers will just gonna think about everything from their or the company's angle, like how to make money.*"

Laws & Regulations. Except for P10, participants generally agreed on the necessity of enacting laws and regulations for the use of ChatGPT, especially considering its potential to generate deceptive information that could mislead users. Although P10 held a different view, their contention primarily revolved around the perceived urgency of implementing such regulations, which differed from others. P10 advocated for a wait-and-see approach, believing that, based on their and their associates' experiences, the use of ChatGPT was solely for entertainment purposes and thus did not necessitate elevation to a legal level. Moreover, concerns are raised about the rapid pace of technological advancement outstripping current legal frameworks, the government's slow response, and the potential for regulations to be either too intrusive or outdated [P4, P10-12]: *"the technology is moving very fast and current measures and in place not really um up to date or quick enough to protect people"*. Simultaneously, participants [P3, P4, P10, P12] acknowledged the complexity inherent in achieving comprehensive global regulation of generative AI such as ChatGPT, given the backdrop of their escalating globalisation. They

pointed out that *"not every country has the same level of expectation from the government, don't have the same safety expectations, etc" [P3]*, and *"I think we sometimes see that already with social media excuse me social media sites so we watch potentially legally in the US or another country might not be here and where does that quite stand and where is the date to hell." [P10]*. However, making some foundational rules that could be adapted and enforced locally could be the initial step. Additionally, as P11 mentioned *"the government, they're not really experts in technology and AI" [P11]*, participants underscored the significance of fostering a collaborative regulatory approach that involves both governmental bodies and industry stakeholders and suggested that industry buy-in could accelerate the process of establishing relevant guidelines.

5 Discussion and Summary of Results

In this section, we outline the major findings related to our research questions. Note that our findings reflect the perceptions of deception among users who have encountered these behaviours. While these insights help clarify how deception is experienced in real-world interactions, they may not be fully generalizable to the broader population.å

- Survey participants primarily reported encountering *over-simplified* responses whereas interview participants identified a new prevalent error: the *partially correct* response. Furthermore, while the survey identified *research* as the primary context for perceived deception, interview data refined this to *idea generation*. (RQ1)
- Survey participants' *perceived responsibility* for deception in ChatGPT and their *behavioural response* are influenced by a combination of personal and other factors. Conversely, *chat-worthiness* perceptions are solely impacted by the *perceived frequency of deception and use*. (RQ2)
- Interview participants expressed varied degrees of *chat-worthiness* in using ChatGPT for both personal and professional purposes. Opinions on *responsibility* for ChatGPT's deployment diverged, with some attributing it to the hosting platform and others emphasising the developers' critical role. Additionally, *behavioural response* to deception differed, with some reducing their usage and others maintaining it, influenced by individual perceptions such as their *pre-existing domain knowledge of ChatGPT*, and *trust*. (RQ2)
- Our chi-square analysis of the survey responses indicates that the *perceived frequency of deception* significantly impacts users' *perceived worthiness of ChatGPT*, independent of personal factors. *Responsibility* for deception is influenced solely by *perceived frequency of deception*. In contrast, *behavioural responses* are shaped by a combination of demographic variables (*gender, age*) and other factors, including *knowledge, verification tendency, and perceived chat-worthiness*. (RQ3)
- Four interview respondents reported a decrease in trust towards ChatGPT post-deception, which illuminates the intricate nature of trust as a construct influenced by more than just deception encounters. The finding also identified additional factors impacting trust in ChatGPT, including *explanatory transparency*, elaborated upon in §4.2.3. (RQ4)
- Interview participants expressed expectations for both general and technical enhancements in ChatGPT to mitigate deception, particularly emphasising the potential of improved *explanations*. A significant portion advocated for user-centric approaches, suggesting that future design and development should more thoroughly incorporate user needs and feedback. Despite acknowledging potential challenges, participants also engaged in a meaningful discussion on the implementation of laws and regulations as viable strategies. (RQ5)

We proceed to discuss the results of our study in more detail.

5.1 Integrated Impact of Mental Models, Experiences, and Personal Factors

The results from the second part of our study focused on interviewing participants (Study II), and highlighted how participants' mental models, experiences (as reported in §4.2.1), and personal factors influence their perception of ChatGPT's deception, and how it affects their trust and changes in their behaviour.

For instance, individuals who do not impose high expectations and standards on ChatGPT for fulfilling their specific requirements or for application in serious contexts tend to maintain a generally relaxed disposition throughout the interview process. As presented in §4.2.2, in areas deemed trivial by the users, there is often no effort to verify the accuracy of information provided by ChatGPT.

Furthermore, there is a tendency among these individuals to blindly trust the information furnished by ChatGPT, often motivated by personal curiosity and the influence of social conformity. In contrast, individuals who approach ChatGPT with the intention of obtaining meaningful and valuable information for task execution exhibit a markedly different interaction pattern. These participants seem to engage in a deeper level of critical thinking concerning the credibility of ChatGPT. The participants (i) actively contemplate the underlying reasons for any discrepancies or errors that they encounter, (ii) deliberate on who might be accountable for these issues, and (iii) consider potential measures for resolution.

The rigorous approach of utilising ChatGPT in this way reflects what can the behaviours observed in privacy-related research. This suggests that there might be an underlying stratification among users based on their concerns and understanding of privacy. For example, prior studies (Hsu et al. 2022; Chen et al. 2023; Soumelidou and Tsohou 2021) have identified that while some individuals are indifferent to the sharing or access of their information by others online, those who prioritise privacy as a paramount concern demonstrate a heightened awareness and demand for privacy safeguards.

Our study found that participants reported varying levels of knowledge about ChatGPT, ranging from moderate to expert, as detailed in §4.2.1 and supported by data in Table 3. However, when asked more detailed questions, we discovered that many participants could not provide thorough or accurate explanations about how ChatGPT works or why ChatGPT might give deceptive answers. This indicates a gap between what participants think they know about ChatGPT and their actual understanding. This knowledge gap challenges the reliability of some of our study's findings, including those from the Chi-square test.

Moreover, our study suggests that people's understanding of ChatGPT's deception is linked to their beliefs about responsibility and knowledge. Specifically, some participants believe that users themselves are responsible for the deceptive responses given by ChatGPT. They argue that the quality of ChatGPT's replies depends on the quality of the information or questions users provide. If the inputs are poor or misleading, ChatGPT's responses will be sub-optimal, suggesting that better user input could improve ChatGPT's performance.

Overall, our study elucidates that personal factors influencing user interaction with ChatGPT extend beyond mere demographics to encompass professional affiliations, job-related experiences, and personal traits that participants disclosed during interviews. For instance, some users described themselves as inherently trusting individuals, while others, particularly those with professional ties to the AI industry or who have witnessed numerous AI failures, expressed a heightened sense of caution towards using ChatGPT. These observations underscore the complexity of user engagement with AI tools and highlight the necessity for a user-centric approach in this domain. We argue that adopting such an approach is not only critical but also in urgent demand within the field, necessitating significant attention from corporations.

5.2 Users' Expectations of Verification and Responsibility

We observed that there seems to be a notable discrepancy between the anticipated convenience offered by the utilisation of ChatGPT and the actual experiences of users. This discrepancy emerges from the findings detailed in §4.2.2, §4.2.3, and §4.2.4. These results highlight participants' concerns regarding the accuracy of

information provided by ChatGPT. Participants found themselves compelled to employ additional measures and exert significant effort to verify the correctness of the information provided by ChatGPT, which seems to contradict the initial rationale behind employing ChatGPT, i.e. the rationale that ChatGPT will make their job easier. As participants have indicated, this verification process often results in a recursive loop where assistance sought from ChatGPT ends up being cross-checked with Google or other databases, raising the question: *why not directly utilise Google or a reliable search engine in the first place?*

Despite this challenge, users acknowledge the benefits of ChatGPT's ability to synthesise and aggregate information, particularly in the early stages of research or exploratory tasks. This trade-off between convenience and the need for verification underscores a critical debate about responsibility in managing deceptive outputs. While some users recognise their role in assessing the accuracy of ChatGPT's responses, they overwhelmingly prefer external verification mechanisms rather than assuming full responsibility themselves. Participants expressed a preference for safeguards such as built-in fact-checking tools or third-party verification systems. However, they remain cautious about completely outsourcing verification, seeking a balance between personal oversight and external support. This reflects an intriguing contradiction: while users demand more reliable AI-generated information, they also expect developers or regulatory bodies to bear the primary responsibility for ensuring accuracy.

5.3 Bridging the Gap: Enhancing Users' Ability to Identify and Manage Deception

While it is true that users must take responsibility for distinguishing between deception and accurate information, they cannot be left to navigate this challenge alone. Our findings suggest that users, despite their awareness of deception, require additional support in improving their ability to critically engage with AI-generated content. This necessitates the development of mechanisms that enhance user capacity while maintaining a balance between individual agency and external intervention.

One effective approach is the integration of AI literacy initiatives that educate users on the limitations and potential biases of ChatGPT. Providing accessible resources, such as interactive tutorials or embedded guidance, could help users develop a more nuanced understanding of how ChatGPT operates, why deception occurs, and how to mitigate its impact.

Secondly, incorporating real-time feedback mechanisms that alert users to potential misinformation could serve as an essential safeguard. For example, confidence scores, source attributions, or fact-checking prompts embedded within ChatGPT's interface could encourage users to critically evaluate responses without imposing excessive cognitive burdens.

In summary, the transition towards user-centric methodologies in the development of ChatGPT and similar LLMs is imperative for ethical, responsible AI development. This shift emphasises the balance between innovation and ethical responsibility, ensuring the creation of technologies that are not only advanced but also safe and aligned with societal values. Our study highlights the urgency of this transition, advocating for a development paradigm that equally values user insights and ethical standards.

5.4 Limitations

Our research is based on self-reported data, which may possibly overlook and not fully capture the complexity of users' experiences and perceptions.

Moreover, the rapid evolution of AI technologies like ChatGPT means that user perceptions and the platform's capabilities could change, potentially dating our findings. Future research should consider longitudinal studies to better track changing user perceptions and AI advancements. Expanding investigations to assess the effectiveness of regulatory and educational measures against deception, as well as developing user-friendly AI verification tools, are crucial next steps.

5.5 Future Directions for Talking Machines

In this paper we addressed the problem of machines, like current LLM-based chatbots, which despite the fact that they are not capable of deliberate deception, they act as vessels of 'banal' deception if placed in particular contexts to interact with humans (Natale 2021; Sarkadi 2023b). So, where does this lead us regarding the development of better communicative AI agents?

We are in the AI age of '*incomplete minds*' to which we increasingly delegate mental tasks to machines (Lewis and Sarkadi 2024). This is also the case regarding the task of communicating or engaging in dialogue with others. A subtle thread throughout this paper is that of *chat-worthiness* of interaction. In the particular context of communicative AI agents, this means the worthiness of talking to such an AI agent. Hence, a pertinent question to ask is the one Charles Hamblin hinted at, namely *How do we build a machine worth talking to?*. To do this, Hamblin proposed the mathematical modelling of dialogue (Hamblin 1971), and finally the design of an AI agent architecture that enables the modelling in a similar fashion of the mind of the AI agent's interlocutor (Staines 2018). This architecture implies several abilities, namely self-awareness, reflection, and Theory of Mind, i.e. the ability to mentalise.

What is the actual **research challenge** in AI here? When AI agents 'talk' about things in a human-interpretable manner they need to make sense, not talk nonsense, blabber, or bullshit. How do we go from AI agents that are designed to fake conversations with us (Walsh 2023) to ones that can engage in deliberate and meaningful conversations, and even use 'genuine' deception in a pro-social manner?

First, work on speech-acts and agent communication languages has to be further developed to enable agents to extract and refer to linguistic semantics from their abstract models of the world (Cohen and Levesque 1995). Second, methods based on natural language processing and argumentation (Cabrio and Villata 2012; Lawrence and Reed 2020) need to be developed for agents to be able to form abstract concepts from linguistic or other types of data. Finally, there is a need to integrate dialogue-based argumentation frameworks (P. McBurney and Parsons 2009) for agents to be able to form sound and consistent arguments, and even tell meaningful stories when interacting with others, without resorting to pre-defined scripts (Schank and Abelson 1975) or by generating output based on statistical patterns. This brings us to the importance of world modelling.

Reflection, as a cognitive process in AI agents, can enable the abstract modelling and re-modelling of the world and others to give semantics to their utterances or even to their non-linguistic behaviour, similarly to the dialogical agents proposed by P. J. McBurney (2002), or the deceptive AI agents proposed by Sarkadi (2021), which have internal 'consequence engines' that simulate the outcomes of communicative interactions with respect to the false beliefs formed in the minds of their interlocutor agents. These sorts of agents not only have the ability to model other agents behaviourally, as explored in the special issue edited by Albrecht and Stone (2018), but have the ability to use an Artificial Theory of Mind to reason about consequences of their actions on the minds of others. For instance, the agents proposed by Sarkadi (2021) use a combination of simulated theory of mind (ST) and theory-theory of mind (TT) to reason about how they can cause changes in the beliefs of others and reason about the consequences of these belief changes, albeit on a high-level for the purpose of human interpretability. Similarly, Winfield's robots use it to predict the actions of other agents and anticipate the likely consequences of those actions both for themselves and the other agents (Winfield 2018). Most importantly, as pointed out by Isaac and Bridewell (2017), when it comes to distinguishing between malicious and pro-social deception, AI agents must be able to reflect and reason about the ethical values of their interlocutors, and at least try to align themselves to those ethical values.

Metacognition, and, especially Theory of Mind, has recently become a very pertinent area of AI research. In particular, neurosymbolic approaches that aim to merge LLM technologies, with symbolic inference engines, inverse planning, abductive reasoning, and cognitive agent architectures (Abrini et al. 2025) along with human-machine argumentation (Trajano et al. 2024), and computational frameworks that model ToM reasoning for

Hybrid Intelligence (Erdogan et al. 2025) seem to be the most promising avenues of developing the capabilities of what Hamblin considered a machine 'worth talking to'.

6 Conclusion

In this study, we aimed to address the issue of 'banal' deception in human-LLM interactions, i.e. how humans perceive deceptive information provided by AI agents, such as ChatGPT, that are not capable of deliberate deception, in various contexts. To do this, we ran two studies into the types of deceptive information encountered by users and the contexts of their occurrence. We highlighted the critical impact of deceptive behaviours on user trust and the varied responses individuals exhibit towards perceived deceptive information. Notably, our findings re-emphasise the need pointed out by Castelfranchi and Tan (2002) more than 20 years ago, namely that we need a multi-dimensional socio-cognitive approach to address both trust and deception in human-AI societies. According to our study, this approach should nowadays consider user education, technical improvements, and robust regulatory frameworks. Furthermore, the study calls for continued exploration into user-centric methodologies and the development of ethical AI systems that prioritise user welfare. Our work aims to set a foundation for *Deceptive AI Ecosystems* by navigating the trade-offs between AI convenience and the need for verification, in order to inform the creation of more transparent, accountable, and trustworthy AI technologies.

Yet, as part of our Deceptive AI Ecosystems approach, these technologies remain just one element of a bigger problem in designing trustworthy AI systems. We should probably change how we talk about AI and deception. An important factor in Deceptive AI Ecosystems, that enhances the deceptiveness of AI technologies, is the creation of context around it and the reinforcement of biases by using AI as a speech act with the ulterior goal of monetisation of individuals' attention (Michel and Gandon 2024) and of society in general (Lewis, Marsh, and Pitt 2021). The trend of democratic backsliding is enhanced not only by the use of technologies but also by the way in which we communicate and 'normalise' our perceptions and beliefs about these technologies, which in today's techno ecosystem is done through social influence by groups of actors/agents, such as BigTech, who have both the power and incentive to do so (Mertzani and Pitt 2022). It is at this level where intentional deception happens, rather than at the technological 'stochastic parroting' level. We believe Coeckelbergh (2018) is right in the sense that deceptive AI is not just about the technology, but as Zhan, Xu, and Sarkadi (2023) points out, it's also about the ecosystem, and we must always keep this in mind when evaluating Deceptive AI technologies. LLMs do not have an incentive to deceive, but improperly regulated human-led organisations do, especially when the global cultural market paradigm promotes market competition at the expense of human values.

As shown by Sarkadi and Lewis (2024), competitive contexts provide the ideal ecosystem for deceptive behaviour to become evolutionarily stable. Looking at this phenomenon from a Cybernetic (Wiener 2019) and Techno-Political (Pitt 2021) perspective of human-AI ecosystems, we can notice that the large-scale deployment of AI technologies and the LLM arms-race[18] between Big Tech will lead, if it has not done so already, to an optimal ecosystem for deceptive behaviour. Hence, our best hope as a truth-seeking AI community in this competitive technological context is to take the role of investigators in the mentalisation arms race against agents of malicious deception (Sarkadi 2023a) whilst looking for ways to promote pro-social AI deception (Castelfranchi 2000).

Supplementary Material: All study materials, including survey questions, interview protocols, and codebooks, are publicly available in the OSF repository https://osf.io/c7upq/.

Acknowledgments

XZ and YX formulated the research questions, designed the study, ran the experiments and performed the statistical analysis. ŞS formulated the research questions, designed, and supervised the study. All authors contributed to the study design and the writing of the paper. We extend our gratitude to our participants for their time and

[18] https://www.economist.com/leaders/2024/05/16/big-techs-capex-splurge-may-be-irrationally-exuberant

invaluable insights. Additionally, we thank the anonymous reviewers and Dr Ruba Abu-Salma for their thoughtful and constructive feedback. ȘS was supported by the Royal Academy of Engineering and the Office of the Chief Science Adviser for National Security under the UK ICRF fellowship.

References

Abrini, Mouad et al. (2025). "Proceedings of 1st Workshop on Advancing Artificial Intelligence through Theory of Mind". In: *arXiv preprint arXiv:2505.03770*.

Albrecht, Stefano V and Peter Stone (2018). "Autonomous agents modelling other agents: A comprehensive survey and open problems". In: *Artificial Intelligence* 258, pp. 66–95.

Alon, Nitay et al. (2023). "A (dis-) information theory of revealed and unrevealed preferences: emerging deception and skepticism via theory of mind". In: *Open Mind* 7, pp. 608–624.

Boden, Margaret A (2016). *AI: Its Nature and Future*. Oxford University Press.

Borji, Ali (2023). "A categorical archive of ChatGPT failures". In: *arXiv preprint arXiv:2302.03494*.

Braun, Virginia and Victoria Clarke (2006). "Using thematic analysis in psychology". In: *Qualitative research in psychology* 3.2, pp. 77–101.

Cabrio, Elena and Serena Villata (2012). "Natural language arguments: A combined approach". In: *ECAI 2012*. IOS Press, pp. 205–210.

Castelfranchi, Cristiano (2000). "Artificial liars: Why computers will (necessarily) deceive us and each other". In: *Ethics and Information Technology* 2.2, pp. 113–119.

Castelfranchi, Cristiano and Yao-Hua Tan (2001). *Trust and deception in virtual societies*. Springer.

– (2002). "The role of trust and deception in virtual societies". In: *International Journal of Electronic Commerce* 6.3, pp. 55–70.

Chakraborti, Tathagata and Subbarao Kambhampati (2019). "(When) Can AI Bots Lie?" In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 53–59.

Chen, Subai et al. (2023). "Research on the influence mechanism of privacy invasion experiences with privacy protection intentions in social media contexts: Regulatory focus as the moderator". In: *Frontiers in Psychology* 13, p. 1031592.

Coeckelbergh, Mark (2018). "How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn". In: *Ethics and Information Technology* 20.2, pp. 71–85.

– (2020). *AI ethics*. MIT press.

Cohen, Philip R and Hector J Levesque (1995). "Communicative Actions for Artificial Agents." In: *ICMAS*. Vol. 95. Citeseer, pp. 65–72.

Delobelle, Jérôme et al. (2020). "Sifting the Arguments in Fake News to Boost a Disinformation Analysis Tool". In: *4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020)*.

Dragan, Anca D, Rachel M Holladay, and Siddhartha S Srinivasa (2014). "An Analysis of Deceptive Robot Motion." In: *Robotics: science and systems*. Citeseer, p. 10.

Ehsan, Upol et al. (2021). "Expanding explainability: Towards social transparency in ai systems". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Epley, Nicholas, Adam Waytz, and John T Cacioppo (2007). "On seeing human: a three-factor theory of anthropomorphism." In: *Psychological review* 114.4, p. 864.

Erdogan, Emre et al. (2025). "TOMA: computational theory of mind with abstractions for hybrid intelligence". In: *Journal of Artificial Intelligence Research* 82, pp. 285–311.

Fleiss, Joseph L, Bruce Levin, and Myunghee Cho Paik (2013). *Statistical methods for rates and proportions*. john wiley & sons.

Frankfurt, Harry G (2005). *On bullshit*. Princeton University Press.
Goffredo, Pierpaolo, Elena Cabrio, et al. (2023). "Disputool 2.0: A modular architecture for multi-layer argumentative analysis of political debates". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 13, pp. 16431–16433.
Goffredo, Pierpaolo, Mariana Espinoza, et al. (2023). "Argument-based Detection and Classification of Fallacies in Political Debates". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 11101–11112.
Hamblin, Charles L (1971). "Mathematical models of dialogue 1". In: *Theoria* 37.2, pp. 130–155.
Hancock, Peter A et al. (2011). "A meta-analysis of factors affecting trust in human-robot interaction". In: *Human factors* 53.5, pp. 517–527.
Hicks, Michael Townsen, James Humphries, and Joe Slater (2024). "ChatGPT is bullshit". In: *Ethics and Information Technology* 26.38, pp. 1572–8439. DOI: 10.1007/s10676-024-09775-5.
Hsu, Chien-Lung et al. (2022). "Privacy concerns and information sharing: The perspective of the U-Shaped curve". In: *Frontiers in Psychology* 13, p. 771278.
Isaac, AM and Will Bridewell (2017). "Why robots need to deceive (and how)". In: *Robot ethics* 2, pp. 157–172.
Jahan, Nusrath and Johnathan Mell (2024). "Unraveling the Tapestry of Deception and Personality: A Deep Dive into Multi-Issue Human-Agent Negotiation Dynamics". In: *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 916–925.
Johnson-Laird, Philip N (2010). "Mental models and human reasoning". In: *Proceedings of the National Academy of Sciences* 107.43, pp. 18243–18250.
Knott, Eleanor et al. (2022). "Interviews in the social sciences". In: *Nature Reviews Methods Primers* 2.1, p. 73.
Kohnke, Lucas, Benjamin Luke Moorhouse, and Di Zou (2023). "ChatGPT for language teaching and learning". In: *Relc Journal* 54.2, pp. 537–550.
Kopp, Tobias, Marco Baumgartner, and Steffen Kinkel (2023). ""It's not Paul, it's a robot": The impact of linguistic framing and the evolution of trust and distrust in a collaborative robot during a human-robot interaction". In: *International Journal of Human-Computer Studies* 178, p. 103095.
Lawrence, John and Chris Reed (2020). "Argument mining: A survey". In: *Computational Linguistics* 45.4, pp. 765–818.
Lee, John D and Katrina A See (2004). "Trust in automation: Designing for appropriate reliance". In: *Human factors* 46.1, pp. 50–80.
Letheren, Kate et al. (2016). "Individual difference factors related to anthropomorphic tendency". In: *European Journal of Marketing* 50.5/6, pp. 973–1002.
Levine, Timothy R (2019). *Duped: Truth-default theory and the social science of lying and deception*. University Alabama Press.
Lewis, Peter R, Stephen Marsh, and Jeremy Pitt (2021). "AI vs "AI": Synthetic Minds or Speech Acts". In: *IEEE Technology and Society Magazine* 40.2, pp. 6–13.
Lewis, Peter R and Stefan Sarkadi (2024). "Reflective artificial intelligence". In: *Minds and Machines* 34.2, pp. 1–30.
Masters, Peta, Michael Kirley, and Wally Smith (2021). "Extended Goal Recognition: A Planning-Based Model for Strategic Deception". In: *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 871–879.
Masters, Peta and Sebastian Sardina (2017). "Deceptive Path-Planning." In: *IJCAI*, pp. 4368–4375.
Masters, Peta, Wally Smith, et al. (2021). "Characterising deception in AI: A survey". In: *Deceptive AI: First International Workshop, DeceptECAI 2020, Santiago de Compostela, Spain, August 30, 2020 and Second International Workshop, DeceptAI 2021, Montreal, Canada, August 19, 2021, Proceedings 1*. Springer, pp. 3–16.
McBurney, Peter and Simon Parsons (2009). "Dialogue Games for Agent Argumentation". English. In: *Argumentation in Artificial Intelligence*. Ed. by Guillermo Simari and Iyad Rahwan. Springer US, pp. 261–280.

McBurney, Peter John (2002). "Rational interaction". PhD thesis. University of Liverpool.
McGuire, Jack et al. (2023). "The reputational and ethical consequences of deceptive chatbot use". In: *Scientific Reports* 13.1, p. 16246.
McHugh, Mary L (2013). "The chi-square test of independence". In: *Biochemia medica* 23.2, pp. 143–149.
Mell, Johnathan, Gale Lucas, et al. (2020). "The effects of experience on deception in human-agent negotiation". In: *Journal of Artificial Intelligence Research* 68, pp. 633–660.
Mell, Johnathan, Gale M Lucas, and Jonathan Gratch (2018). "Welcome to the real world: How agent strategy increases human willingness to deceive". In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 1250–1257.
Mertzani, Asimina and Jeremy Pitt (2022). "Social influence and the normalization of surveillance capitalism: Legislation for the next generation". In: *IEEE Technology and Society Magazine* 41.2, pp. 57–63.
Michel, Franck and Fabien Gandon (2024). "Pay attention: a call to regulate the attention market and prevent algorithmic emotional governance". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 7. 1, pp. 971–983.
Mittelstadt, Brent Daniel et al. (2016). "The ethics of algorithms: Mapping the debate". In: *Big Data & Society* 3.2, p. 2053951716679679.
Natale, Simone (2021). *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press, USA.
Nolan, Susan A. (2023). *Learning to Lie: The Perils of ChatGPT*. https://www.psychologytoday.com/intl/blog/misinformation-desk/202303/learning-to-lie-the-perils-of-chatgpt. Blog. (Visited on 03/16/2023).
Pacchiardi, Lorenzo et al. (2023). "How to catch an ai liar: Lie detection in black-box llms by asking unrelated questions". In: *arXiv preprint arXiv:2309.15840*.
Patterson, Mathew (Mar. 2024). *Using ChatGPT for Customer Service*. URL: https://www.helpscout.com/blog/chatgpt-customer-service/.
Peer, Eyal, Joachim Vosgerau, and Alessandro Acquisti (2014). "Reputation as a sufficient condition for data quality on Amazon Mechanical Turk". In: *Behavior research methods* 46.4, pp. 1023–1031.
Piazza, Nancirose and Vahid Behzadan (2023). "A Theory of Mind Approach as Test-Time Mitigation Against Emergent Adversarial Communication". In: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 2842–2844.
Pitt, Jeremy (2021). *Self-organising multi-agent systems: Algorithmic foundations of cyber-anarcho-socialism*. World Scientific.
Price, Adrian et al. (2023). "Domain-Independent Deceptive Planning". In: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 95–103.
Ray, Partha Pratim (2023). "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope". In: *Internet of Things and Cyber-Physical Systems*.
Rohrbach, Anna et al. (2018). "Object hallucination in image captioning". In: *arXiv preprint arXiv:1809.02156*.
Sarkadi, Stefan (2021). "Deception". PhD thesis. King's College London.
– (2023a). "An arms race in theory-of-mind: Deception drives the emergence of higher-level theory-of-mind in agent societies". In: *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*. IEEE, pp. 1–10.
– (2023b). "Deceptive AI and Society". In: *IEEE Technology and society magazine* 42.4, pp. 77–86.
– (2024). "Self-Governing Hybrid Societies and Deception". In: *ACM Transactions on Autonomous and Adaptive Systems* 19.2, pp. 1–24.
Sarkadi, Stefan and Peter R Lewis (2024). "The Triangles of Dishonesty: Modelling the Evolution of Lies, Bullshit, and Deception in Agent Societies". In: *Proc. of the 23rd International Conference on Autonomous Agents and*

Multiagent Systems (AAMAS 2024). International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).
Sarkadi, Stefan, Peidong Mei, and Edmond Awad (2023). "Should my agent lie for me? A study on attitudes of US-based participants towards deceptive AI in selected future-of-work scenarios". In: *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).
Sarkadi, Stefan, Alison R. Panisson, et al. (2019). "Modelling Deception using Theory of Mind in Multi-Agent Systems". In: *AI Communications* 32.4, pp. 287–302.
Sarkadi, Stefan, Alex Rutherford, et al. (2021). "The evolution of deception". In: *Royal Society Open Science* 8.9, p. 201032.
Sarkadi, Stefan, Ben Wright, et al. (2021). *DeceptiveAI*. Vol. 1296. Springer.
Schank, Roger C and Robert P Abelson (1975). "Scripts, plans, and knowledge". In: *IJCAI*. Vol. 75. New York, pp. 151–157.
Schneiderman, Ben and Michael Muller (2023). *On AI anthropomorphism. Medium. April 10, 2023*.
Searle, John Rogers (1969). *Speech acts: An essay in the philosophy of language*. Vol. 626. Cambridge university press.
Shanahan, Murray (2024). "Talking about large language models". In: *Communications of the ACM* 67.2, pp. 68–79.
Soumelidou, Aikaterini and Aggeliki Tsohou (2021). "Towards the creation of a profile of the information privacy aware user through a systematic literature review of information privacy awareness". In: *Telematics and Informatics* 61, p. 101592.
Staines, Phillip (2018). *Linguistics and the Parts of the Mind: Or how to Build a Machine Worth Talking to*. Cambridge Scholars Publishing.
Such, Jose et al. (2017). "Photo privacy conflicts in social media: A large-scale empirical study". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 3821–3832.
Tandoc Jr, Edson C, Zheng Wei Lim, and Richard Ling (2018). "Defining "fake news" A typology of scholarly definitions". In: *Digital journalism* 6.2, pp. 137–153.
Tiffany, Hsu and Thompson Stuart A. (2023). *Disinformation Researchers Raise Alarms About A.I. Chatbots*. https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html. Blog. (Visited on 03/18/2023).
Trajano, Guilherme et al. (2024). "Translating natural language arguments to computational arguments using llms". In: *Computational Models of Argument*. IOS Press, pp. 289–300.
Tripathi, Salil (2023). *We asked ChatGPT about its impact on human rights and business. Here's what it told us*. Blog.
Tversky, Amos and Daniel Kahneman (1988). "Rational choice and the framing of decisions". In: *Decision making: Descriptive, normative, and prescriptive interactions*, pp. 167–192.
– (1996). "On the reality of cognitive illusions". In: *Psychological Review* 103.3, pp. 582–591.
Valmeekam, Karthik et al. (2023). "On the planning abilities of large language models-a critical investigation". In: *Advances in Neural Information Processing Systems* 36, pp. 75993–76005.
Verma, Mudit, Siddhant Bhambri, and Subbarao Kambhampati (2024). "Theory of Mind abilities of Large Language Models in Human-Robot Interaction: An Illusion?" In: *arXiv preprint arXiv:2401.05302*.
Vock, Ido (2022). *ChatGPT proves that AI still has a racism problem*. https://www.newstatesman.com/quickfire/2022/12/chatgpt-shows-ai-racism-problem. Blog. (Visited on 03/15/2023).
Walsh, Toby (2023). *Faking it: Artificial intelligence in a human world*. La Trobe University Press.
Wardle, Claire and Hossein Derakhshan (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Vol. 27. Council of Europe Strasbourg.
Wiener, Norbert (2019). *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press.

Winfield, Alan FT (2018). "Experiments in artificial theory of mind: From safety to story-telling". In: *Frontiers in Robotics and AI* 5, p. 75.

Xiao, Yangyu and Yuying Zhi (2023). "An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions". In: *Languages* 8.3, p. 212.

Xiao, Yijun and William Yang Wang (2021). "On hallucination and predictive uncertainty in conditional language generation". In: *arXiv preprint arXiv:2103.15025*.

Zhan, Xiao, Yifan Xu, and Stefan Sarkadi (2023). "Deceptive AI Ecosystems: The Case of ChatGPT". In: *Conversational User Interfaces, CUI'23, July 19–21, 2023, Eindhoven, Netherlands*. ACM.

Received 28 March 2025; accepted 03 July 2025