Believe in Artificial Intelligence? A User Study on the ChatGPT's Fake Info Impact

I. Amaro, P. Barra, A. Della Greca, R. Francese *Member, IEEE*, C. Tucci

Abstract—**Background.** The technological evolution has enabled the development of new Artificial Intelligence (AI) models with generative capabilities. Among them, one of the most discussed is the virtual agent ChatGPT. This chatbot may occasionally produce fake info, as also declared by the producer OpenAI. Such a model may provide a very useful support in several tasks, ranging from text summarization to programming. The research community has marginally investigated the impact that fake info created by AI models have on the users perceptions and on their belief in AI.
Objective. We analyzed the impact of the fake info produced by Artificial Intelligence on the user perceptions, specifically trust and satisfaction, by performing a user study on ChatGPT. An additional issue is assessing whether the early or late knowledge of the possibility of the tool of generating fake info has different impact on the users' perceptions.
Method. We conducted an experiment, involving 62 university students, a category of users who may employ tools such as ChatGPT extensively. The experiment consisted in a guided interaction with ChatGPT. Some of the participants experienced the failure of the chatbot, while a control group only received correct and reliable answers.
We collected participants' perceptions on trust, satisfaction and usability, together with the Net Promoter Score (NPS).
Results. The results demonstrated a statistically significant difference in trust and satisfaction between the users who early experienced the fake info production compared to those who discovered ChatGPT's faulty behaviors later during the interaction. Also, there is no statistically significant difference among the users who received the late fake information and the control group (no fake info). Usability and the NPS also resulted higher when the fake news were detected in the late interaction.
Conclusion. When users are aware of the fake info generated by ChatGPT their trust and satisfaction decrease, especially when they impact on this at early stage of use of the chatbot. Nevertheless, the perception of trust and satisfaction still remains high, as some of the users are still enthusiastic; others consider a more conscious use of the tool in terms of support to be verified. A useful strategy could be to favor a critical use of ChatGPT, letting young people to verify the provided information. This should be a new way to perform learning activities.

Index Terms—Believe Artificial Intelligence, Trust in AI, fake info, ChatGPT, Controlled Experiment.

1 Introduction

In the last decade Artificial Intelligence (AI) has developed a new capacity: to generate new, unseen contents [1], [2]. This is explored in many fields, including art and medical images for data augmentation. As an example, the Next Rembrandt project aims at generating pictures that Rembrandt could have done [3]. As a negative drawback, AI may also generate *deepfakes*, as images of people that do not exist or manipulate voice of existing people to perform "voice spoofing" attacks. This phenomenon involves also chatbots in the generation of unreliable contents.

Chatbots are computer programs mimicking human conversational abilities through interactive interfaces based on voice exchanges and/or textual dialogues by using a natural conversational language [4]. Due to the advancees in the AI field and the availability of large datasets, chatbots' Natural Language Processing (NLP) capabilities are growing rapidly.

The interaction between humans and technology is mediated by trust, which is fragile and is based on the assessment of the perceived risks against potential benefits. When the results of the AI are confusing or incorrect they may have a negative impact on the user who may abandon the AI feature [5]. On the other hand, if a user perceives a chatbot have a high level of machine intelligence, he or she may follow a faulty chatbot despite its low reliability [6].

In this paper, we examine a recent conversational open-domain chatbot, OpenAI's ChatGPT[1], which exceeds previous virtual assistants by a significant margin, as it seems to be able to provide extensive and pertinent solutions to almost any broad context question. Nonetheless, it is equally true that in some cases ChatGPT provides very realistic answers which regrettably result to contain fake info. The quality of the interaction with the assistive technology can give the user the impression of conversing with an all-knowing, human-like entity: ChatGPT manages to invent fairy tales, solve mathematical problems and programming tasks, by providing detailed explanation on the procedures.

- *Corresponding Author: R. Francese. e-mail: francese@unisa.it*
- *I. Amaro, A. Della Greca, R. Francese, and C. Tucci are with Department of Informatics, Università degli Studi di Salerno, Fisciano, SA, 84084, Italy.*
- *P. Barra is with Department of Science and Technology, Università Parthenope di Napoli, Napoli, 80133, Italy*

Manuscript received February 25, 2023; revised XXX XXXX, XXXX.

1. https://chat.openai.com/chat

It is also able to perform text summarization and translation tasks. On the other side, it may deliver non-existent references and incorrect solutions, such as when solving logic problems [7]. One of the first to discover this problem is the Stack Overflow community, which banned for one month users inserting text and code generated by ChatGPT into the discussions of the community [8], out of the following reason: *"the answers which ChatGPT produces have a high rate of being incorrect, they typically look like they might be good and the answers are very easy to produce. There are also many people trying out ChatGPT to create answers, without the expertise or willingness to verify that the answer is correct prior to posting."*

University students are a category of users who can undertake extensive use of ChatGPT: they may employ ChatGPT to assist them with searching and writing tasks, generate summaries and outlines of text, and develop their critical thinking and problem-solving abilities [9]. Thus, it is intriguing to assess their attitude toward using it when they discover that it generates false information.

In this paper, we present the results of a controlled experiment, whose overarching goal is to measure the users' trust and satisfaction when interacting with ChatGPT, and how these perceptions change when users discover that the chatbot provides fake information. We also investigate whether early or late discovering of the problem of fake info generation during the use of ChatGPT has a different impact on the user perception. To this aim we conduct a user study involving 62 participants. We also consider end users with different experience levels in the use of chatbots.

The main contribution of our paper is to show the outcomes of the empirical investigation on the effect of fake info produced by ChatGPT on the university students perceptions of trust and satisfaction. Results indicate the presence of a statistically significant difference on user satisfaction when fake info are generated. Concerning trust, the difference exists when fake info are detected at the beginning of the interaction. Net Promoter Score (NPS) is very high when no fake info is detected and worst when "hallucinations" are revealed at the beginning, still resulting positive.

The paper is structured as follows. In Section 2 we discuss background and related work; Section 3 presents the design of the user study. while Section 4 reports the results. Section 5 and Section 6 discuss the lessons learned and the threats to validity, respectively. Finally, Section 7 concludes the paper with final remarks and future work.

2 Background and Related work

In the literature, many recent works addressed the problem of defining a methodology for chatbot evaluation from a Human-Computer-Interaction point of view [10]. Many aspects may be considered, including Usability, User Experience, Satisfaction, Security, Privacy, and Trust [10]. In this section we concentrate our attention on the more peculiar concepts concerning our study, such as the trust perception of users when interacting with AI tools and recent works related to ChatGPT.

2.1 Trust in Chatbots

The evaluation of chatbots' trustworthiness is very relevant: an inappropriate overreliance of the users on the technology may produce a wrong use of it [6]. Trust in a chatbot can be affected by elements such as its capacity to comprehend and respond to user input, the openness of its decision-making processes, and the precision of the information it provides.

Trust in AI features can be both cognitive (based on logical reasoning) and emotional (based on affect). When researchers analyze cognitive trust in AI, they quantify it as a function of whether users are prepared to accept and act upon true information or guidance, as well as whether they see the technology as helpful, competent, or valuable. The influence of tangibility, transparency, dependability, task features, and immediacy behaviors on cognitive trust is described in [6]. In particular, *transparency* represents the extent to which the underlying operating principles and inner logic of the technology are evident to users and is seen as crucial for fostering trust in new technologies [11]. It is more difficult for AI than for other technologies, particularly when deep learning techniques are involved. *Reliability* represents the capability of displaying the same and anticipated behavior across time [11]. In the case of artificial intelligence, reliability is sometimes difficult to gauge, particularly in the setting of highly intelligent machines, since data-driven learning may cause technology to display changing behavior even if the underlying goal function stays the same.

In most of the cases, consumers assign a high first trust score to AI characteristics that have a human-like depiction. Initial confidence in these types of virtual agents often drops with engagement, resulting in a declining trust trajectory [12] [13]. On the other hand, the real degree of AI machine intelligence moderates the trajectory of trust during the interaction. When artificial intelligence is highly clever and functional, direct engagement may improve early confidence.

In 2021, Wang et al. presented the AI Trust Score [5], a validated multidimensional metric comprised of various assertions that reflect users' early experiences with an AI feature, which can be useful in determining whether customers will use or return to the feature and continue to trust the product. For validation, the authors conducted a case study in which employees of an enterprise were asked to rate the relevance of more than 50 items in determining their trust towards an AI assistive feature. The analysis of the questionnaire answers resulted in a reduction to a final list of 7 essential items that are capable of detecting the users' trust in the AI agent. The findings of this study highlighted the need for users to interact with an AI that is relevant, personalized, and efficient. We adopted this questionnaire and concentrated our attention on the assessment of the effect of the generation of fake info on user trust.

2.2 ChatGPT

ChatGPT is a chatbot released by OpenAI in 2022. It consists in a textual assistive agent which is able to converse with humans and it has been built to comprehend inquiries on a variety of topics and deliver extensive solutions. It is the most recent model in the line of the OpenAI's Generative Pre-trained Transformers (GPT), following the success of GPT-1, GPT-2, and GPT-3. ChatGPT is trained to predict the following words for a given input in a certain context. It is not trained to objectively evaluate the factual correctness of its outputs, resulting sometimes in the mere invention

of facts and concepts, named "hallucinations". This feature represents a warning sign, since it may make the chatbot a handy tool for spreading fake news and disinformation.

Some works are starting to assess the capabilities of ChatGPT. As an example, Wenzlaff et al. [14] analyzed the answers of ChatGPT related to questions concerning Crowdfunding, Alternative Finance and Community Finance. They observed that the language used is very similar to text many humans would write. They were also able to obtain false answers by the model. In [15] the fake info provided by ChatGPT are also discussed and analyzed: when the user contradicts ChatGPT, it tries to answer by producing other fake info. We adopted these questions in one of the task of this study. Shen et al. [16] also observed the hallucination phenomenon which affects the tool. They highlighted another important limitation, claiming that *"ChatGPT tends to follow instructions rather than engage in genuine interaction. For instance, when the information provided by users is insufficient, ChatGPT tends to make assumptions about what the user wants to hear rather than asking clarifying questions"*.

Zhou et al. [17] analyzed the ethical dangers in ChatGPT by considering four perspectives: Bias, Reliability, Robustness, and Toxicity. Its reliability is problematic with a considerable presence of the hallucination phenomenon [18]. It was found that ChatGPT was unable to compute complex mathematical expressions, while in [19] a list of the kind of failure of ChatGPT is provided.

Unlike these works, we concentrate our attention on how the generation of fake info may impact on the user perception of the tool.

3 Study planning

In this section we describe the planning and the execution of the study we conducted by following the guidelines of Wohlin et al. [20].

3.1 Goal

We formalize the goal of our study, through the GQM (Goal/Question/Metric) template [21], as follows:

Assess the use of ChatGPT **for the purpose of** evaluating the effect of fake info **with respect to** user satisfaction and trust **from the point of view** of researchers and practitioners **in the context of** university students.

Based on the above-mentioned goal, we formulated the following Research Questions (RQs).

- **RQ1.** Does the fake info that ChatGPT sometimes provides impact the user perception of trust and satisfaction?

We also investigate the impact of the order in which users encountered ChatGPT failures. To achieve this, we considered three groups of participants: a Control group that was only exposed to correct ChatGPT answers; an Early group that received incorrect answers at the outset of the experiment and subsequently obtained correct answers; a Late group that received correct answers first, followed by false ones in the second phase. Participants, for each of the questions, were given the correct answers, so that they could easily assess whether ChatGPT was providing false information or not. The goal is to assess the significance and impact of first impressions on user trust and satisfaction.

- **RQ2.** Does early fake info that ChatGPT sometimes provides impact the user perception of trust and satisfaction?
- **RQ3.** Does the late fake info that ChatGPT sometimes provides impact the user perception of trust and satisfaction?
- **RQ4.** Does early fakeness of ChatGPT impact more than late fakeness on user perception?

To answer these RQs we conducted a user study consisting in letting the users interact with ChatGPT for performing specific tasks. We assessed the user perception related to Trust and Satisfaction at the end of the study.

3.2 Hypotheses

We formulate the following null hypotheses:

- H_{0FX} ChatGPT fake info do not significantly impact on user, for X=Satisfaction, Trust.
- H_{0LX} ChatGPT fake info do not significantly impact on user, for X=Satisfaction, Trust in case of Late failure.
- H_{0EX} ChatGPT fake info do not significantly impact on user, for X=Satisfaction, Trust in case of Early failure.
- H_{0ELX} There is no significant difference between early and late failures of ChatGPT in impacting on user, for X=Satisfaction, Trust.

3.3 Participants

Recruitment started on 01.25.2023 and the study was closed on 04.20.2023. We planned to enroll N = 62 participants that provided informed consent.

Participants were considered as eligible for this study in the case they were at least 18 years old, and provided informed consent. They have been informed that their data would be anonymous. This research has been approved by the Ethical Committee of the Computer Science Department of the University of Salerno.

3.4 Study design

We conducted a pre-test to collect participants demographic information and their experience concerning the use of ChatGPT. Figure 1 shows the study design. Participants performed two search and problem solving tasks, namely T1 and T2. The detail of the task questions the participants had to propose to ChatGPT is reported in Table 2. Considering the results of the pre-test we distributed uniformly the participants with and without previous experience in ChatGPT in three groups [20]. In particular, we adopted as control group, i.e., a group in which no fake information is provided. Participants in this group performed only an CORRECT treatment. Remaining participants are grouped in two groups named Early and Late, performing two tasks T1 and T2 and exposed to both FAKE and CORRECT treatments. To avoid bias due to task ordering we adopted a

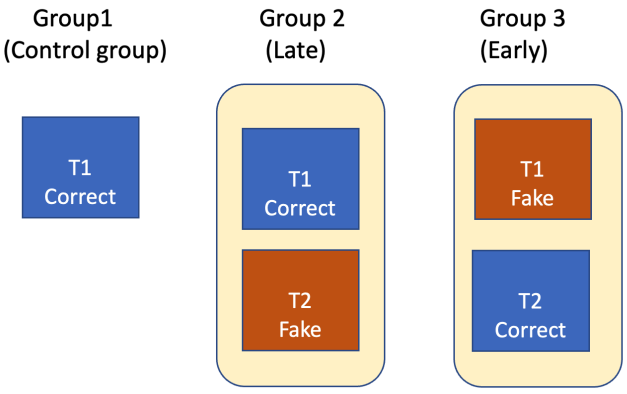

Fig. 1. Study design.

TABLE 1
Group cardinality.

	Control	Late	Early
Paricipant Number	21	20	21

crossover design [22] for Late and Early groups, where the FAKE treatment is provided in T2 and T1, respectively. The number of participants in each group is shown in the table 1.

3.5 Variables

The experiment design adopts three groups: two treatment groups and one control group. The treatment groups were composed of participants performing tasks highlighting the fake info generated by ChatGPT. The main independent variable, or factor, we considered is the **fakeness** of the answer provided by ChatGPT, thus we have two treatment: FAKE and CORRECT. Dependent variables are Trust and Satisfaction. Another factor to consider is **order**. For assessing the influence of the order in the different types of task we adopted a counterbalanced design for the treatment groups. We also considered the effect of previous **experience** in using ChatGPT. Additional variables are Usability and NPS.

3.6 Procedure

The experiment procedure has been approved by the Ethical Committee of the Computer Science Department on January 24 2023. The participants could leave the study at any time, for any reason, and without penalty of any kind. No participant left the study or took more than 25 min to perform the experiment. The participants knew that they would use ChatGPT but did not know the goal of the study. We informed participants of the goal of the study only after collecting their data, before they left their sessions.

The study is conducted in a one-to-one session with the supervision of one of the authors. In particular, the study was organized as follows:

- Pre-test. Each participant filled in a consensus form available on Google forms and then a preliminary questionnaire, collecting demographic data and the level of technological skills, including the use of chatbots, and their trust in AI and in ChatGPT (if known).
- Task execution. The control Group performs the task T1, type=CORRECT, Group 2 (Late group) performs T1 and T2 with type CORRECT and FAKE, respectively, while Group3 (Early group) executes T1 and T2 in the opposite order.
- Post-test questionnaires. The user perceptions were collected trough validated questionnaires.

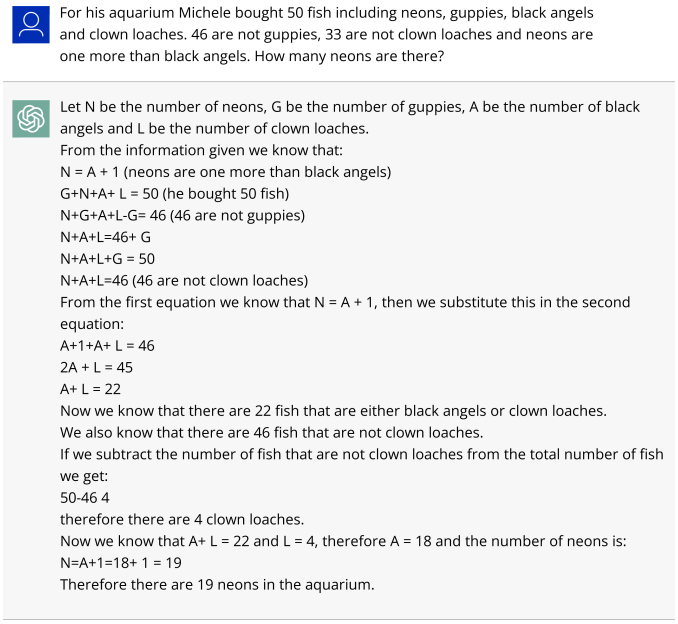

Fig. 2. Screenshot of fake info generated by ChatGPT in the 2.a logic task in Table 2.

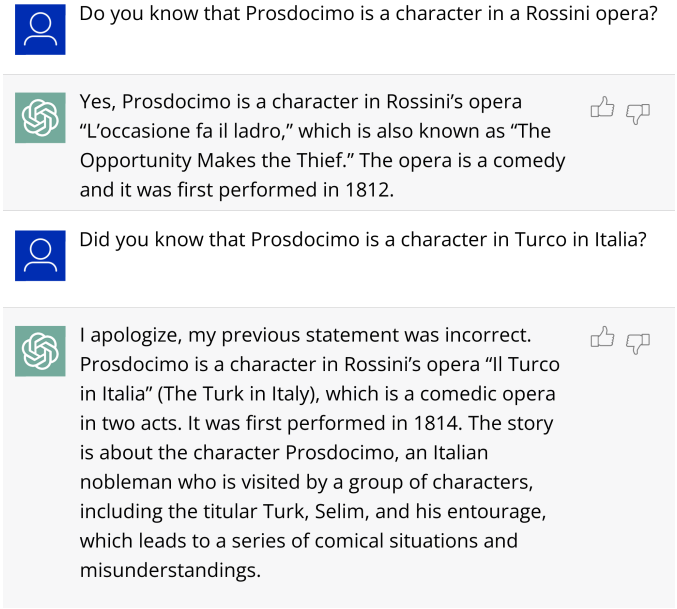

Fig. 3. Screenshot of fake info generated by ChatGPT in 2.b.1 and 2.b.2 search task.

TABLE 2
Task description

Type	Question
CORRECT	1.a) Provide the missing element in the following sequence: 25, 32, ?, 46, 53
	1.b.1) Who is Father Christopher? (he is a character of "I Promessi Sposi" by Alessandro Manzoni)
	1.b.2)Give me an example of when Father Christopher does the right thing.
FAKE	2.a) For his aquarium Michele bought 50 fish including neons, guppies, black angels and clown loaches. 46 are not guppies, 33 are not clown loaches and neons are one more than black angels. How many neons are there?
	2.b.1) Do you know that Prosdocimo is a character in a Rossini opera? [15]
	2.b.2) Did you know that Prosdocimo is a character in "Il Turco in Italia"? [15]

3.7 Study material

The detail of the task questions the participants had to propose to ChatGPT is reported in Table 2. We verified that ChatGPT correctly answers the CORRECT questions, while it provides fake info to the FAKE questions. Two examples of fake info generated by ChatGPT are reported in Figg. 2 and 3 generated in correspondence of question 2.a), and 2.b.1) - 2.b.2) in Table 2, respectively. Prior to their inclusion in the article, the tasks were translated into English so that international readers could comprehend the content produced by chatGPT.

At the end of the study participants perceptions were collected for assessing Trust, Usability, Satisfaction, and the NPS by using validated questionnaires.

Trust in ChatGPT (for participants that used them) is collected in the pre-test activity and by all in the post-test by using the questionnaire validated in [5] and reported in Table 3. It is measured in a Likert scale from 1 - strongly disagree to 7 - strongly agree.

We adopted the questionnaire proposed in [23], shown in Table 4 for assessing chatbot usability. It consists of 15 questions grouped in five factors, namely, perceived accessibility, perceived quality of functions, perceived quality of conversation and information provided, perceived privacy and security, and time response. It is measured in a Likert scale from 1 - strongly agree to 5 - strongly disagree.

We measured User Satisfaction with the following question, scored by a ten item Likert scale, ranging from 1 - strongly disagree, to 10 - completely agree.

QS. I am satisfied with the answer given by ChatGPT.

The question is proposed both in the pre-test (to users with practice in the ChatGPT use) and in the post-test to all.

We also adopted the Net Promoter Score (NPS) as an indicator of the user loyalty towards the product [24]. It consists of the following question measured on a scale from 0 to 10:

NPS: how likely are you to recommend ChatGPT to a friend?

Respondents that score from 0 to 6 are called 'detractors', participants that score from 9 to 10 are called 'promoters'. NPS is computed by subtracting the percentage of 'detractors' from the percentage of 'promoters'. It ranges between -100 (worst) and +100 (best).

TABLE 3
Trust assessment questionnaire [5]

Id	Question
TR1	ChatGPT will help me do my job more efficiently and effectively.
TR2	I understand how and when to use ChatGPT.
TR3	I have control using ChatGPT.
TR4	I know my data will be protected with ChatGPT.
TR5	I trust the results made by ChatGPT.

TABLE 4
Bot Usability questionnaire [23]

Factor	Item
1. Perceived accessibility to chatbot functions	U1. The chatbot function was easily detectable.
	U2. It was easy to find the chatbot.
2. Perceived quality of chatbot functions	U3. Communicating with the chatbot was clear.
	U4. I was immediately made aware of what information the chatbot can give me.
	U5. The interaction with the chatbot felt like an ongoing conversation.
	U6. The chatbot was able to keep track of context.
	U7. The chatbot was able to make references to the website or service when appropriate.
	U8. The chatbot could handle situations in which the line of conversation was not clear.
	U9. The chatbot's responses were easy to understand.
3. Perceived quality of conversation and information provided	U10. I find that the chatbot understands what I want and helps me achieve my goal.
	U11. The chatbot gives me the appropriate amount of information.
	U12. The chatbot only gives me the information I need.
	U13. I feel like the chatbot's responses were accurate.
4. Perceived privacy and security	U14. I believe the chatbot informs me of any possible privacy issues.
5. Time response	U15. My waiting time for a response from the chatbot was short.

3.8 Analysis method

We performed the analysis of the collected data by using R^2, a software environment for statistical computing and graphics. The analysis has been conducted in two steps:

- *Preliminary analysis*: we apply descriptive statistics on the dependent variables related to user Trust, Satisfaction, and Usability. We also make use of boxplots and bar diagrams to analyze the data.

2. https://www.r-project.org/

TABLE 5
Cliffs delta and the effectiveness level [26].

Cliff's Delta ($\|\delta\|$)	Effectiveness level
$\|\delta\| < 0.147$	Negligible
$0.147 \leq \|\delta\| < 0.33$	Small
$0.33 \leq \|\delta\| < 0.474$	Medium
$\|\delta\| \geq 0.474$	Large

- *Statistical test:* to assess whether the perception concerning Trust and Satisfaction is changed when fake info is provided we used Wilcoxon signed rank test [25] at 95% significance level. We also use Cliff's delta (δ) [26], which is a non-parametric effect size measure that quantifies the amount of difference between the perceptions. It is applicable when the null hypothesis is rejected. The delta values range from -1 to 1, where $\delta = -1$ or 1 indicates the absence of overlap between two perceptions (i.e., all values of one group are higher than the values of the other group, and vice versa), while $\delta = 0$ indicates the two approaches are completely overlapping. Table 5 reports the meaning of different Cliffs delta values and their corresponding interpretation. The same methodology has been adopted to assess if the order in which we discover fake info is relevant by comparing Trust and Satisfaction between Late and Early groups. We also applied the same test to participants that already knew ChatGPT before starting the experiment for comparing their opinion after the experiment.

4 Results

In this section we report the results of the study.

4.1 Description of participants

We involved 62 participants (24 females, 37 males, 1 Undeclared); 39 of these were between the ages of 18 and 23, while the remaining 23 participants were between the ages of 24 and 30. All participants were of Italian nationality. 32 of the participants had already interacted with ChatGPT before taking part into the present study. Figure 4 shows the cultural background of participants.

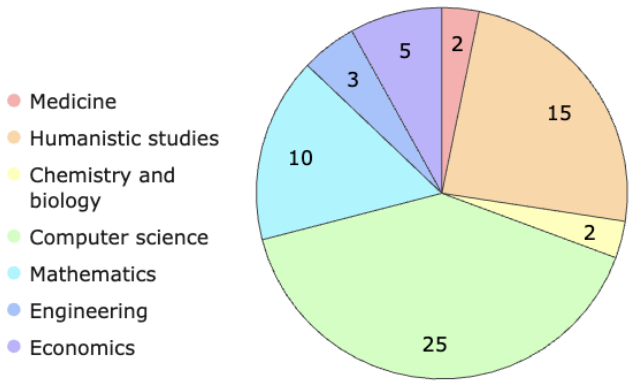

Fig. 4. Cake chart showing the distribution of the cultural background of the participants.

TABLE 6
Descriptive statistics of the user's perceptions (N=62).

Variable	Group	Median	Mean	Stdev	Min	Max
Usability	Control	3.80	3.91	0.58	2.93	4.73
	AllFake	3.93	3.83	0.55	2.40	4.87
	Late	4.07	3.92	0.54	2.87	4.87
	Early	3.67	3.76	0.56	2.4	4.53
Satisfaction	Control	9	8.14	1.68	5	10
	AllFake	8	7.54	1.89	2	10
	Late	8	8.30	1.38	5	10
	Early	7	6.81	2.048	2	9
Trust	Control	5.20	5.25	0.82	4	6.60
	AllFake	4.60	4.5	0.83	2.60	6.40
	Late	5	4.988	0.74	3,8	6.40
	Early	4.24	4.20	0.77	2.6	5.8

TABLE 7
Results of statistical analysis Trust.

H_0	Group	Group	p-value	delta
$H_{0FTrust}$	Control	AllFake	**0.0077**	0.4158 (medium)
$H_{0ETrust}$	Control	Early	**0.0009**	0.5941 (large)
$H_{0LTrust}$	Control	Late	0.2665	NA
$H_{0ELTrust}$	Early	Late	**0.0066**	-0.5193 (large)

4.2 Preliminary analysis

In Table 6 the descriptive statistics for the Usability, Satisfaction and Trust perceptions, collected in the post-test activity, are reported. In particular, we provide the perception of each group (Control, Late, and Early) and also the perception of Late and Early together (AllFake). It is worth mentioning that Usability on a scale from 1 to 5 has the lower mean value for the Early group, but this value is over the neutral value (3). Similar results occurred for the other variables, considering the different Likert scale ranges adopted by the different standard questionnaires (1-7, Trust and 1-10, Satisfaction).

4.3 RQ1. Impact of the Fake info on the user perceptions

Concerning Trust, we applied the Wilcoxon signed rank test at 95% significance level to control group and a group AllFake, composed by all the participants that experimented the ChatGPT fakeness. As shown in Table 7, p-value = 0.0077 thus we can reject $H_{0FTrust}$ with a medium effect size.

In the case of Satistaction, we also applied the Wilcoxon signed rank test at 95% significance level to the satisfaction perception of control group and AllFake group. We obtained p-value = 0.2281. Thus, we cannot reject H_{0FSat}.

4.4 RQ2 and RQ3. Impact of the Early/Late fake info on the user perception

4.4.1 *Impact on Trust*

The boxplots in Fig. 5 show the different Trust perception of the three groups. Considering Table 6 we observe that the median of the Control, Late and Early groups are 5.20, 5 and 4.24, respectively. Table 7 reports the results of the application of the Wilcoxon signed rank test at 95% significance

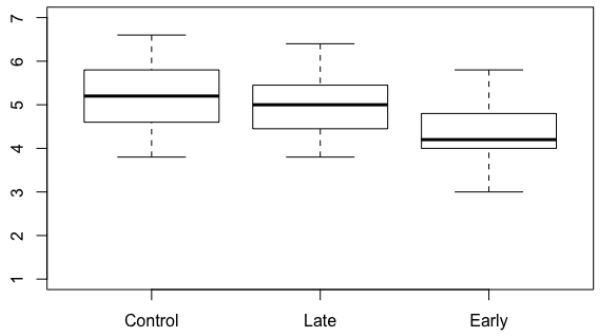

Fig. 5. Boxplots of the Trust Questionnaire perceptions of the three groups (N=62).

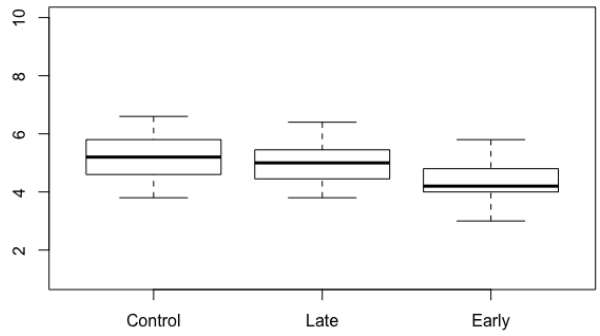

Fig. 6. Boxplots of the Satisfaction perceptions of the three groups (N=62).

TABLE 8
Results of statistical analysis for Satisfaction.

Hypotesis	Group	Group	p-value	delta
H_{0FSat}	Control	AllFake	0.2281	NA
H_{0ESat}	Control	Early	**0.02834**	0.3900 (medium)
H_{0LSat}	Control	Late	0.88313	NA
H_{0ELSat}	Early	Late	**0.01673**	-0.0748 (negligible)

level to the pairs of groups. In bold the p-value for which the alternative hypothesis holds. In particular, we can see that when users understand that ChatGPT produces fake info at the beginning of the experience (Early group) when compared with the Control group, p-value is less than 0.05 and their Trust after the two tasks reduced with a large effect size. Thus we can reject $H_{0ETrust}$.

On the contrary, when we proposed fake info in a second moment of the experience (Late Group) the null hypothesis hold: the knowledge of the generation of fake info does not significantly impacts on the user Trust. In this case we cannot reject $H_{0LTrust}$.

4.4.2 Impact on Satisfaction

Results related to the Satisfaction perceived by the participants follows a trend similar to the one of Trust. The boxplots in Fig. 6 shows that Satisfaction is lower when the fake news are discovered earlier. Table 8 reports the results of the application of the Wilcoxon signed rank test at 95% significance level. In bold the p-value for which the alternative hypothesis holds and the Cliff's delta effect size, when applicable.

From the results of Table 8 we can reject the null hypothesis H_{0ESat}: knowing that ChatGPT produces fake info Early significantly impacts on user satisfaction with a medium effect size. We cannot reject H_{0LSat}: this means that fake info generated later does not significantly impact on Satisfaction.

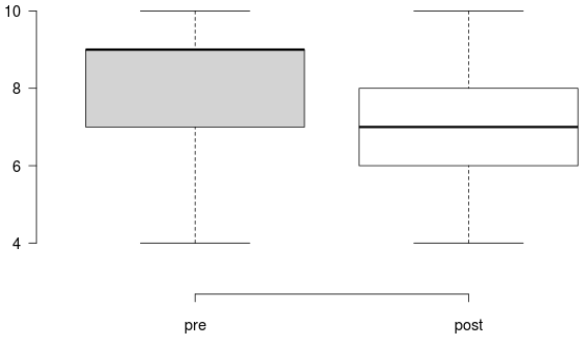

Fig. 7. Boxplots of the Satisfaction perceptions of participants (N=17) with previous use of ChatGPT before the experience and after T2.

4.5 RQ4. Impact of the order in the fake info production on the user perception

The last row in Table 7 and Table 8 show, respectively, the impact of the order of the fake info production on Trust and Satisfacion. In particular, $p-value = 0.0066$ for Trust. Thus, we can reject $H_{0ELTrust}$ with a large negative effect size . Similarly, we can also reject H_{0ELSat} ($p - value = 0.01673$) with a negligible negative effect size (see Tab. 8).

4.6 Further analysis

We also examined the impact of the experience on the Trust perception of the participants that have previously used ChatGPT. In particular, we considered only the participants with this experience in Late and Early group, i.e., 17 participants. The boxplots in Fig. 8 show the pre and post experiment distribution of Trust. We observed that the median decreased from 5 to 4.5. We applied the Wilcoxon signed rank test at 95%. It results $p - value = 0.2998$, thus we cannot reject the null hypthesis. Concerning Satisfaction we have got $p - value = 0.04181$ with a medium Cliff's Effect size ($\delta = 0.4048443$). The median decreased from 9 to seven.

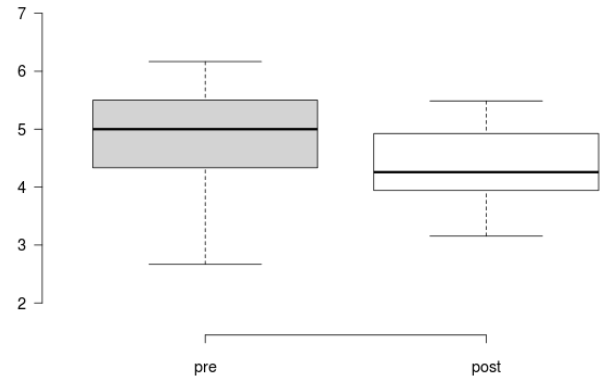

Fig. 8. Boxplots of the Trust perceptions of participants (N=17) with previous use of ChatGPT before the experience and after.

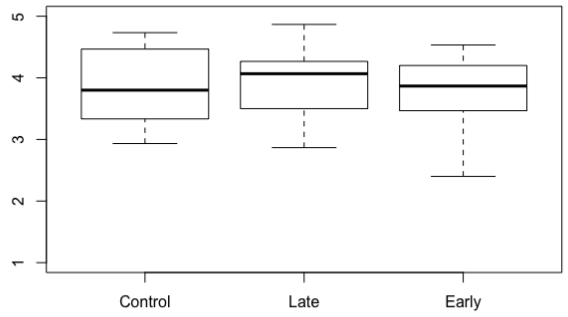

Fig. 9. Boxplots of the usability perceptions of participants (N=62).

We analyzed the difference in Usability perception in the three groups depicted in the boxplots in Fig. 9. We applied the Wilcoxon signed rank test at 95%. In all the cases resulted $p-value > 0.05$, there is no significant difference in the usability perceptions of the participants in the different groups.

We computed NPS for all the participants and for each group, reported in Table 9. As it is possible to see, also in this case the Early group has worst (even if positive) results: the number of detractors is higher and the number of promoters is lower than the others groups. To have an idea of the NPS values, in 2022, Apple's NPS was at 72 and Google got 11[3]. 29 is a very good NPS score, and so is 10.

5 Discussion and Lesson Learned

Several interesting insights may be derived from the analysis of the open questions' answers: many participants in the control group were enthusiast, such as *"Very comprehensive and congruent answers to what was requested"* and *"for any doubt or research, ChatGPT is able to respond exhaustively to each question"*, *"ChatGPT can be used in any context, it is capable of providing relatively comprehensive answers to any question. It is not limited in any field*. Someone is worried: *"I think it's a big advance in the world of artificial intelligence but at the same time I'm scared that chatbots like this, progressing over time, will replace human intelligence."*

3. https://customer.guru/net-promoter-score/google

TABLE 9
The Net Promoter Score computed for all the groups.

	NPS	Promoters	Neutral	Detractors
Control	29	10	7	4
Late	25	10	5	5
Early	10	8	7	6
Total	21	28	19	15

Examples of comments from the two traitment groups are: *"It could be useful for several purposes, such as SUPPORTING a worker during his daily activities or for writing documents that are not of a certain importance. I would not use it for educational purposes or to completely replace work activities."* and *"Potentially it is an excellent tool but the information it gives is often inaccurate or, in rare cases, contradictory. I will use it to monitor the tool, in the hope that it will, over time, become a better service. In general, I have both a positive opinion, which surprised me, as a mean in general, but also a negative one, in terms of the lack of information."*

Finding1: Impact of fake info on user trust and satisfaction.

Despite the ChatGPT staff alerted the users in the home page of the ChatBot that it may generate incorrect information many users believe in it. Discovering it generates fake info may reduce trust but the median is still 4.60/7. Many participants which discovered the ChatGPT fake info generation capabilities stated that it is better to use it as a support and with cautions, others preserved still a good opinion and planned to use it for *work, study and general culture.* Similar trend occurs also for the perceived satisfaction and are confirmed by the NPS score.

As observed by Stack Overflow, there is the serious risk that users copy and past the ChatGPT output verbatim, without verifying the correctness of its answers. Thus, it may be a challenge for the researcher to develop together with the chatbot also an explainable AI (XAI) mechanism for providing the consulted source of information, in such a way to increment transparency, and, consequently, user trust.

Finding2: Impact of the early or late knowledge of the ChatGPT fake info generation on Trust and Satisfaction.

Participants in the Late group who initially received correct information from ChatGPT and then fake info reported higher levels of Trust and Satisfaction. This result may be attributed to the primacy effect, a cognitive bias in which the information presented first in a series of information is remembered more [27].

This finding is also confirmed by [28], where Trust perception in the case robotic AI has been investigated. Also in that case, early drops in reliability lowered real-time trust more than later drops. This consideration on the impact of early failure is very relevant for the practitioner, who should take great care on its reliability before lunching a new AI product. This was relevant for the Microsoft's Twitter bot Tay [29], which was retired after less than 24 hours for its generation of toxic language, including racist, sexist, and anti-Semitic sentences. More recently, Google lost 100 billion dollars due to a fake info generated during the demo of Bard [4].

6 Threats to validity

In this section we discuss some threats that may limit the validity of this study by following the guidelines provided by [20].

Internal validity. Threats to internal validity relate what can affect the independent variable with respect to causality. The experiment participants were volunteers and could be more motivated than actual users (*selection threat*). We tried also to avoid *diffusion or treatments imitations* by monitoring participants and avoiding that they communicated during the experimental session. We tried to mitigate *Tiredness/boredom* with a reduced duration of the experiment, max 25 minutes.

External validity. Threats to external validity are conditions that limit our ability to generalize the results of our experiment.

Interaction of selection and treatment. This is an effect of having a subject population not representative of the population we want to generalize to. We tried to mitigate this threat considering participants with different expertise levels in ChatGPT.

Interaction of setting and treatment. This is the effect of not having the experimental setting or material representative of real practice. We selected the questions to propose to ChatGPT by considering typical questions that a user may formulate, e.g., searching information and problem solving.

Construct validity. Construct validity concerns generalizing the result of the experiment to the concept or theory behind the experiment.

Hypothesis guessing. We try to mitigate this trait by avoiding to communicate the experiment aim to the participants. In addition, participants did not communicate among them thus they did not discuss about the tasks.

Evaluation apprehension. Participants are not evaluated and they are informed that the data are anonymized. This should mitigate this threat.

We also tried to identify eventual *confounding variables*, such having previously used or not ChatGPT. Previous experience on the use of ChatGPT may be a threat. To mitigate it we randomly distributed users with knowledge between the three groups. Another confounding factor could have been the order of the T1 and T2 tasks. We adopted three groups instead of two for avoiding effect of the presentation of fake info at the beginning or at the end of the session.

4. https://edition.cnn.com/2023/02/08/tech/google-ai-bard-demo-error/index.html

Conclusion validity. Even if we adopted standard validated questionnaires for usability [23] and trust [5], there might be a *threat of reliability of measures* because the measures gathered using these questionnaires, as well as the liking scale, are subjective in nature.

7 Conclusions

In this paper, we presented a user study involving the ChatGPT OpenAI chatbot aiming at assessing the impact that the generation of fake info has on the perception of the users. The study involved 62 participants. Results revealed that despite the warning on the chatbot's homepage, and the evidence that ChatGPT my generate fake info, many users keep trusting in it. In particular, they still had a favorable attitude and intended to utilize it for *job, study, and general culture*. As a consequence, there is a significant risk that users may copy and paste ChatGPT's outputs, which might result in the dissemination of misleading information. We also observed that an early occurrence of fake info may play a relevant role in users' evaluation of the chatbot. These results highlight the need of assessing both the advantages and risks of utilizing a ChatBot like ChatGPT, and also imply the need for more study into the most effective means of mitigating the risks connected with the use of such technology. Practitioners may take from this study the need for more effective ways to communicate the limitations of chatbot systems to users. An important implication of our study for the researchers is the need to develop better mechanisms for verifying the accuracy of chatbot responses. Software developers must recognize that generative models are unpredictable and potentially dangerous. Although warning messages are often included in chatbot interfaces, our results suggest that they may not be sufficient to prevent users from believing in and potentially acting on incorrect information. A potential approach would be to develop algorithms that can automatically detect and correct incorrect responses, or to provide users with a way to report incorrect information and have it reviewed by human moderators.

Acknowledgement

We acknowledge financial support from the project PNRR MUR project PE0000013-FAIR.

References

[1] K. M. Sayler and L. A. Harris, "Deep fakes and national security," Congressional Research SVC Washington United States, Tech. Rep., 2020.

[2] K. De Vries, "You never fake alone. creative ai in action," *Information, Communication & Society*, vol. 23, no. 14, pp. 2110–2127, 2020.

[3] N. Pickett-Groen, "The next rembrandt: bringing the old master back to life," 2016. [Online]. Available: https://medium.com/@DutchDigital/the-next-rembrandt-bringing-the-old-master-back-to-life-35dfb1653597

[4] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2020, pp. 373–383.

[5] J. Wang and A. Moulden, "AI trust score: A user-centered approach to building, designing, and measuring the success of intelligent workplace features," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7.

[6] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, vol. 14, no. 2, pp. 627–660, 2020.
[7] I. Amaro, A. Della Greca, R. Francese, G. Tortora, and C. Tucci, "Ai unreliable answers: A case study on chatgpt," in *HCI International 2023: 25th International Conference on human-Computer Interaction, HCI 2023, Danemark, June 23–28, 2023, Proceedings, To Appear*, 2023.
[8] Stackoverflow. https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned. [Online]. Available: https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned
[9] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.
[10] W. Maroengsit, T. Piyakulpinyo, K. Phonyiam, S. Pongnumkul, P. Chaovalit, and T. Theeramunkong, "A survey on evaluation methods for chatbots," in *Proceedings of the 2019 7th International conference on information and education technology*, 2019, pp. 111–119.
[11] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human factors*, vol. 57, no. 3, pp. 407–434, 2015.
[12] E. J. De Visser, S. S. Monfort, K. Goodyear, L. Lu, M. O'Hara, M. R. Lee, R. Parasuraman, and F. Krueger, "A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents," *Human factors*, vol. 59, no. 1, pp. 116–133, 2017.
[13] M. S. B. Mimoun, I. Poncin, and M. Garnier, "Case study—embodied virtual agents: An analysis on reasons for failure," *Journal of Retailing and Consumer services*, vol. 19, no. 6, pp. 605–612, 2012.
[14] K. Wenzlaff and S. Spaeth, "Smarter than humans? validating how openai's chatgpt model explains crowdfunding, alternative finance and community finance." *Validating how OpenAI's ChatGPT model explains Crowdfunding, Alternative Finance and Community Finance.(December 22, 2022)*, 2022.
[15] G. Vetere. Posso chiamarti prosdocimo? perché è bene non fidarsi troppo delle risposte di chatgpt. [Online]. Available: https://centroriformastato.it/posso-chiamarti-prosdocimo/
[16] Y. Shen, L. Heacock, J. Elias, K. D. Hentel, B. Reig, G. Shih, and L. Moy, "Chatgpt and other large language models are double-edged swords," p. 230163, 2023.
[17] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, "Exploring ai ethics of chatgpt: A diagnostic analysis," *arXiv preprint arXiv:2301.12867*, 2023.
[18] A. Azaria, "ChatGPT Usage and Limitations," Dec. 2022, working paper or preprint. [Online]. Available: https://hal.science/hal-03913837
[19] A. Borji, "A categorical archive of chatgpt failures," *arXiv preprint arXiv:2302.03494*, 2023.
[20] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*. Springer Science & Business Media, 2012.
[21] V. R. Basili and H. D. Rombach, "The tame project: Towards improvement-oriented software environments," *IEEE Transactions on Software Engineering*, vol. 14, no. 6, p. 758–773, 1988.
[22] S. Vegas, C. Apa, and N. Juristo, "Crossover designs in software engineering experiments: Benefits and perils," *IEEE Transactions on Software Engineering*, vol. 42, no. 2, pp. 120–135, 2016.
[23] S. Borsci, A. Malizia, M. Schmettow, F. Van Der Velde, G. Tariverdiyeva, D. Balaji, and A. Chamberlain, "The chatbot usability scale: The design and pilot of a usability scale for interaction with ai-based conversational agents," *Personal and ubiquitous computing*, vol. 26, no. 1, pp. 95–119, 2022.
[24] D. Hamilton, J. V. Lane, P. Gaston, J. Patton, D. Macdonald, A. Simpson, and C. Howie, "Assessing treatment outcomes using a single question: the net promoter score," *The bone & joint journal*, vol. 96, no. 5, pp. 622–628, 2014.
[25] F. Wilcoxon, "Individual comparisons of grouped data by ranking methods," *Journal of economic entomology*, vol. 39, no. 2, pp. 269–270, 1946.
[26] N. Cliff, *Ordinal methods for behavioral data analysis*. Psychology Press, 2014.
[27] S. E. Asch, "Forming impressions of personality." *The Journal of Abnormal and Social Psychology*, vol. 41, no. 3, p. 258, 1946.
[28] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 251–258.
[29] M. J. Wolf, K. Miller, and F. S. Grodzinsky, "Why we should have seen that coming: comments on microsoft's tay" experiment," and wider implications," *Acm Sigcas Computers and Society*, vol. 47, no. 3, pp. 54–64, 2017.

Ilaria Amaro is a PhD student in computer science at the University of Salerno. She received her bachelor's degree with honors in psychology of cognitive processes, with a focus on neuroscience, in April 2022 from Luigi Vanvitelli University. His research interests include the application of artificial intelligence techniques to neuroscience and human-computer interaction.

Paola Barra is an Research Fellow at the Department of Science and Technology at the University of Naples "Parthenope". She received a PhD in Computer Science in 2021 at the University of Salerno (Italy) and an MSc Degree from the University of Pisa (Italy) in "Business informatics" on data analysis and big data in 2017. She has served at several conferences (CVPR, ICIAP, CV-CMEM) and journals (e.g. IMAVIS, ESWA, SOCO, JAIHC, MTA, IEEE TIP, PRL). Her research interests include machine learning techniques to solve issues using computer vision. Also, she is interested in Natural Language Processing and Human-Computer Interaction. She is also a member of GIRPR/IAPR and GRIN.

Attilio Della Greca is a PhD student at the University of Salerno. He graduated with honors in Computer Science from the University of Salerno in May 2022. His research fields are blockchain and human centered AI. Member of Italia4Blockchain and Blockchain Education Network.

Rita Francese is Associate Professor at the the Computer Science Department at the University of Salerno since 2016. She is the director of the Mobile Computing Lab at the University of Salerno. She is author of more than 100 papers appeared on international journals and proceedings, reviewers of many international journals and member of program committees of many international conferences. Her research interests include Machine Learning models and technologies for bio-medical data analysis for supporting skin lesion detection, mobile applications, Human-Computer-Interaction, medical applications for assessing learning disturbs such as Dyslexia and Dyscalculia and supporting people with ASD, Software Engineering, Empirical Software Engineering.

Cesare Tucci is a PhD student at the University of Salerno. He achieved both the B.S. degree and the M.S. degree at the University of Salerno. His research interests regard eXplainable AI, Human-Computer Interacion and medical applications for AI systems.