Protecting victim and witness statement: examining the effectiveness of a chatbot that uses artificial intelligence and a cognitive interview

Protecting Victim and Witness Statement: Examining the Effectiveness of a Chatbot that Uses Artificial Intelligence and a Cognitive Interview

Rashid Minhas, Camilla Elphick & Julia Shaw
Dr Rashid Minhas

Senior Lecturer in Criminology and Psychology, School of Human and Social Sciences, University of West London, UK Rashid.Minhas@uwl.ac.uk

Dr Camilla Elphick

Research Associate, Faculty of Arts & Social Sciences, Open University, UK
camilla.elphick@open.ac.uk

Dr Julia Shaw
Research Fellow, Division of Psychology and Language Sciences, University College London, UK. drjuliashaw@gmail.com

***Corresponding author**
Dr. Rashid Minhas
School of Human & Social Sciences
University of West London
St Mary's Road, Ealing, London W5 5RF
Rashid.Minhas@uwl.ac.uk

Funding sources and conflict of interest declaration

Financial support for this project has been obtained from All Turtles, which made all development of the tool, and all research of it possible. The researchers have not been paid for any specific results and have preregistered the study. Still, the team recognises this financial support as a potential source of bias, which is part of the motivation to make the tool widely accessible to all researchers, including those who are not affiliated with All Turtles.

Protecting Victim and Witness Statement: Examining the Effectiveness of a Chatbot that Uses Artificial Intelligence and a Cognitive Interview

Abstract

Information of high evidentiary quality plays a crucial role in forensic investigations. Research shows that information provided by witnesses and victims often provide major leads to an inquiry. As such, statements should be obtained in the shortest possible time following an incident. However, this is not achieved in many incidents due to demands on resources. This intersectional study examined the effectiveness of a chatbot (the AICI), that uses artificial intelligence (AI) and a cognitive interview (CI) to help record statements following an incident. After viewing a sexual harassment video, the present study tested recall accuracy in participants using AICI compared to other tools (i.e., Free Recall, CI Questionnaire, and CI Basic Chatbot). Measuring correct items (including descriptive items) and incorrect items (errors and confabulations), it was found that the AI CI elicited more accurate information than the other tools. The implications on society include AI CI provides an alternative means of effectively and efficiently recording high-quality evidential statements from victims and witnesses.

Keywords: Artificial intelligence; Victim statement, Witness statement, Memory recall, workplace harassment

Introduction

Information of high evidentiary quality plays a decisive role in forensic investigations (Milne and Bull 2016). Obtaining good quality and reliable information is vital for successful investigations (Bull 2013). However, memory is fallible, and people are often unable to recall critical forensic details such as person descriptors (Kebbell and Milne 1998; Read and Connolly 2017), and memory can become distorted over time. Erroneous testimony is recognised as the leading cause in the failure of forensic investigations (Kaplan, Damme, Levine, and Loftus 2016; Rossmo 2016).

The quality of victim and witness accounts is time-critical. A significant problem is a delay between individuals experiencing an event and reporting their account of it (Gabbert, Hope and Fisher 2009). Ideally, they should be given an opportunity to do so as soon as possible, as their memory is also vulnerable to the influence of post-event information (Shaw and Porter 2015). Research suggests that delay and post-event information compromise recall completeness and accuracy (e.g., Gabbert, Memon and Allan 2003; Gabbert et al. 2009; Tuckey and Brewer 2003). For example, the accuracy of a report decreases as the delay between the incident and recall increases (see Read and Connolly 2017; Wixted and Ebbesen 1997). In an experiment to test witness recall, Turtle and Yuille (1994) found that participants recalled approximately 43 per cent fewer details about a simulated incident after a 3-week delay compared with being interviewed immediately.

The Cognitive Interview (CI) is widely used to obtain a detailed memory report from cooperative interviewees. The CI combines principles of cognitive and social psychology (Fisher, Milne, and Bull 2011). In essence, CI is a systematic set of tools that accesses an individual's

memory without altering it and is not hampered by poor phrasing (Milne and Bull 1999). The CI has consistently enhanced the quantity and quality of information obtained from an interviewee (Stein and Memon 2006) and has been found to elicit up to 40 per cent more information from eyewitnesses in comparison to a standard police interview, without decreasing accuracy (Memon, Meissner, and Fraser 2010). If used effectively, it does this without producing more incorrect responses or increasing susceptibility to leading questions (Westera, Kebbell, and Milne 2011). The positive effects of the CI have been well replicated and are robust, but it is not without limitations. One important drawback concerns the demands placed upon resources, mainly due to the length of time taken to conduct a full CI (Gabbert et al. 2009).

Additionally, as part of the cognitive interview, it is important to let people freely talk about what happened, without being interrupted (Milne and Bull 1999). But interviewers find this aspect of CI exceptionally challenging. Studies have found that during one of the most critical parts of the CI, the free-recall phase, people are interrupted every few seconds (Poole and Lamb 1998). Interviewers seem only to be able to last between 2-10 seconds (on average) before interrupting (Fisher and Geiselman 2010).

The present study examines whether an AI chatbot rooted in the CI techniques (AI CI) can offer a reliable and widely accessible way to help people recall and report what happened. Gabbert et al. (2009) found that self-administered cognitive interview (SAI) allows a comprehensive, immediate recall attempt. They found that participants using SAI recalled more correct details in the delayed recall task than control participants (see Gabbert et al. 2009). These authors argue that SAI has the potential to become a widely used recall tool for forensic investigations. An AI chatbot rooted in the CI should also remove social situational demands (e.g. those associated with face-to-face interviews [see Perfect et al. 2008 for more details]). In turn, reducing task-related perceptual

load (see Murphy and Greene 2016) and releasing more cognitive resources for invoking episodic retrieval (Taylor and Dando 2018). Additionally, chatbots are available at any time, which allows individuals to report and record immediately after an incident. This can be a simple and effective way to minimise problems associated with delay and obtain more reliable accounts from victims and witnesses.

The present study examined the effectiveness of AI CI against three other digital tools. In this experiment, participants watched a video involving workplace sexual harassment before using one of the digital tools to document their memory of what happened. These tools were; (i) the AI CI, (ii) a Basic Chat rooted in the CI technique but without NLP; (iii) a Questionnaire based on the CI technique; and (iv) a Free Recall tool with only one open-ended question. All conditions enabled individuals to record their memories in their own words.

We tested whether the AI CI can produce relevant and accurate information than the alternative conditions. We hypothesised that the AI CI would help people to report correct information than the other conditions, as relevant CI probes have been shown via Natural Language Processing (LNP) to help people to provide more information (Fisher and Geiselman 2010). We considered that any increase in incorrect information might be driven by more descriptive details such as person descriptors (Fisher and Geiselman 2010), as the probes would invite people to give more information about items that they had already mentioned. We tentatively hypothesised that the AI CI could reduce the amount of incorrect information elicited, as the NLP probes will shape the response around information that had previously been given (Murphy and Greene 2016) rather than asking general questions that could prompt more biased responses (Minhas et al. 2017). As such, previous research studies (e.g., Fisher and Geiselman 2010) found

that relevant and non-leading probs minimised incorrect information during victim and witness interviews.

Chatbots

In computer-mediated communication (CMC), the use of online human-like cues is viewed as one of the most important developments in online interface applications (Prendinger and Ishizuk 2013). The use of such features also found its way to Facebook in early 2016, revealing an innovative tool: a chatbot. Chatbots are' 'machine conversation systems that interact with human users via natural conversational language'' (Shawar and Atwell 2005, p. 489). The instant responses are usually comprised of structured messages, links, or even specific call-to-action buttons. A variety of new chatbot architectures and technologies (e.g., Ultra Hal, ALICE, Jabberwacky, Cleverbot) have emerged, each endeavouring to mimic natural human language more accurately and meticulously (Rahman, Mamun, and Islam 2017; Shawar and Atwell 2007).

There is potential to extend chatbots to human roles. Sectors such as finance, health, retail, and law are increasingly adopting AI and chatbots into everyday functions (Brandtzaeg and Følstad 2017). Despite much research investigating the efficacy of chatbots in other domains, little is known about whether chatbots can record witness and victim accounts quickly and effectively. As such, the present study examines whether an AI chatbot rooted in CI techniques can offer a reliable and widely accessible way to help people recall and report what happened.

The Present Study

Novel Administration: Implementing CI via an AI chatbot (AI CI)

Anyone can interact with AI CI chatbot, which walks them through the CI and includes follow-up questions. By using AI CI instead of a human, the user can create a time-stamped pdf report contemporaneously, creating high-quality memory evidence that can be used immediately or at a

later date. AI CI also ensures that interview script is always followed, and the user is never interrupted or subject to a conscious/unconscious biased credibility assessment/judgment that could happen with a real human (due to explicit and implicit biases, see Minhas, Walsh and Bull 2017; Greenwald and Krieger 2006).

Well-crafted AI CI is also not prone to suggestibility (Ridley 2013). Suggestibility refers to the extent to which an interviewee will adjust their recall depending on external factors (such as the interviewer's authority or confirming to others' testimony) (Ridley 2013). Suggestibility is generally considered negative within the context of forensic interview, as people can change their answers from those based on their own actual recall to suggested answer, irrespective of accuracy (Lamb et al. 2008; London et al. 2013).

Unlike human interviews, with AI CI, the user can do a memory interview immediately after an incident, from anywhere, in private, and can take as long as they need. They also do not need to log in or register to use it; they can simply open the webpage and begin. Finally, users are told that no human is interacting with them, giving them a real, and perceived, privacy that would not be possible in an in-person interaction.

The Architecture of the AI CI

The general architecture involves three main components; (i) the interface and back-end code that allows interaction with the AI CI; (ii) a mostly linear decision-tree that tells the AI what questions to ask and when; and (iii) the use of trained Natural Language Processing (NLP) to enable the AI to ask relevant follow-up questions.

Interface and back-end code: The front-end code that powers the AI CI was built using React, a JavaScript library. This allows the user to see and use the AI CI in their web-browser. The back-end code is ***Node.js,*** an open-source run-time environment that uses JavaScript. The back-end is

where we process reports, where the AI lives and is responsible for digitally signing the time-stamped report PDF. This cryptographically signed document cannot be changed anymore and can act as crucial evidence. The database we use is called MongoDB, which is an open-source database program. All the above is stored on Amazon Web Services, which allows easily scalable cloud computing.

Decision-trees: the flow of the AI CI was designed to mimic the cognitive interview, which uses an established script and protocol. See Figure 1 for a visualisation of the general AI CI decision tree.

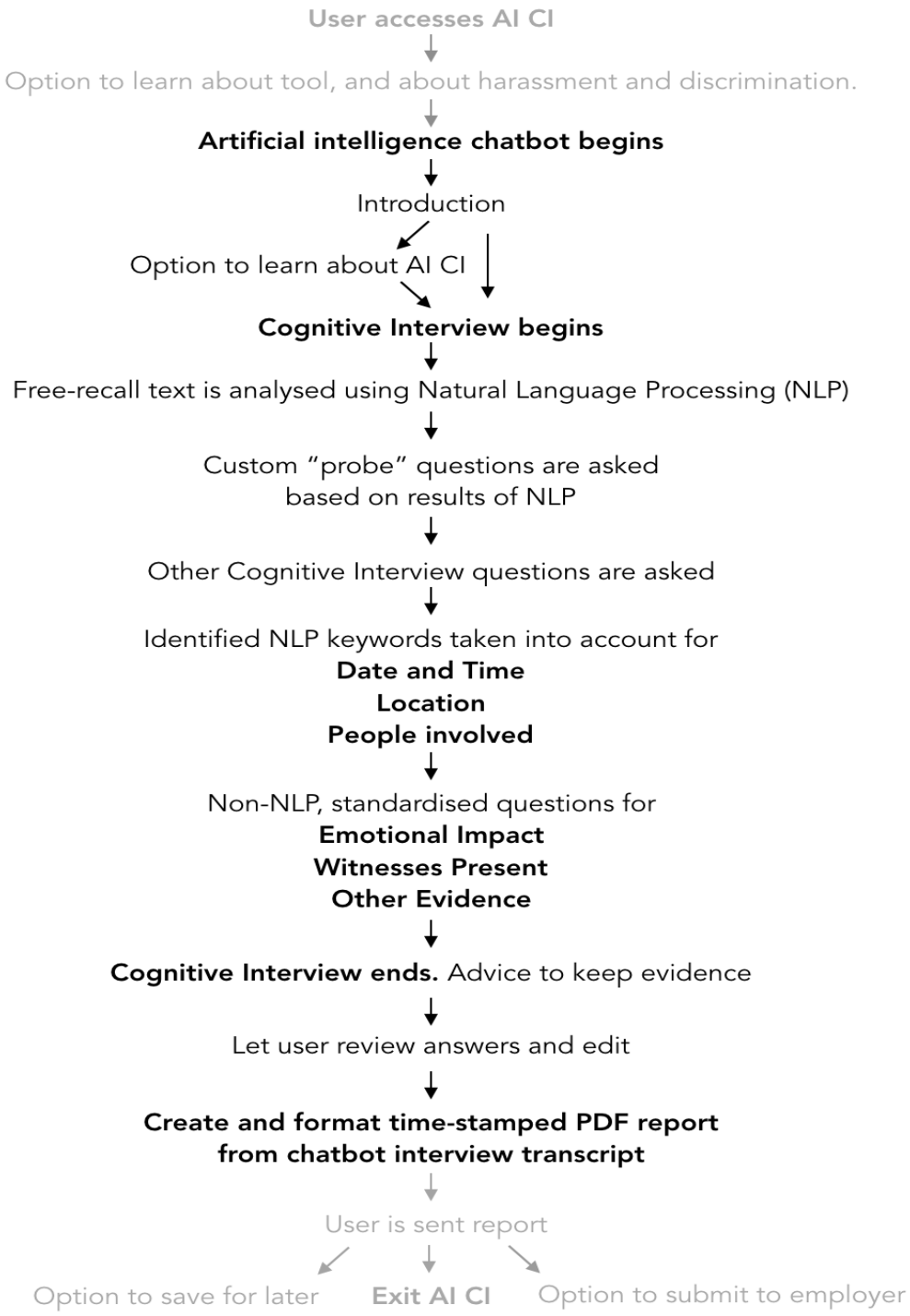


Figure 1. AI CI general decision tree structure

We initially copied the script verbatim, but have edited it over time following feedback that was received during the user testing. It is in line with the original structure of the CI, but sections have been modified for clarity and making the AI CI more user friendly. We used this ability to obtain, revise, and test specific parts of the CI to be a particular benefit of using an online AI tool by the wider public.

Natural Language Processing. The 'intelligent' part of the AI CI is the ability to do Natural Language Processing (NLP). NLP is a well-established technique that has been in development for the past three decades. The AI CI was built using the SpaCy NLP database, which allows the AI to analyse sentences and words to identify keywords, modelling human language. The AI can use context to learn which words are important for follow-up questions. Information on how we trained our AI CI[1] is in Appendix 1.

We tested whether the AI CI can elicit accurate and comprehensive accounts from an interviewee than the alternative conditions. We hypothesised that the AI CI would help people to report more correct information than the other conditions, as relevant CI probes have been shown to help people to provide more information (Fisher and Geiselman 2010). We considered that any increase incorrect information might be driven by more descriptive details such as person descriptors (Fisher and Geiselman 2010), as the probes would invite people to give more information about items that they had already mentioned. We tentatively hypothesised that the AI CI could reduce the amount of incorrect information elicited, as the NLP probes would shape the response around information that had previously been given (Murphy and Greene 2016) rather

[1] **AI CI for Research:** The AI CI has tremendous potential to study the effectiveness of the CI in different contexts, and widely accessible. For the purposes of this research, we created a research version of the AI CI, and the data presented in the present study were collected using this research version. This research version of the AI CI is available to all who want to use it for research purposes.

than asking general questions that can produce more biased responses (Minhas et al. 2017). As such, research studies (e.g., Fisher and Geiselman 2010) found that the non-leading probes minimised incorrect information during victims and witness interviews (Fisher and Geiselman 2010).

Methods

Participants

One hundred and sixty-one adults from the general population took part in the present study. Thirty-three of these participants were excluded for either; (i) not understanding the task; or (ii) for writing only one sentence to the first question. One of the participants was excluded for reporting that he was intoxicated while doing the task. This left 127 participants (63 males, 52 females, one non-binary, and 11 who gave no response) of which 77 were Caucasian, seven were Afro-Caribbean, six were Asian, three were Mixed, one was Arab, one was Latin-American, and 32 participants chose not to disclose their ethnicity. Participants' ages ranged from nineteen to 54 years (M = 29.68, SD = 7.60).

Materials

The stimuli consisted of one audio-video of a sexual harassment scenario that lasted 1 minute and 45 seconds. In the video, a man arrives in a woman's office and praises her for her work. He then invites her on a date, hinting that this would be in exchange for putting in a good word for her promotion. Despite her declining and appearing visibly uncomfortable, he persists until eventually giving up and leaving the room, stating that she apparently was not interested in the promotion. The video started with some information about the woman and how long she had been in the company, and with the date, time, and locations in text form. After the scenario, the video ended with additional information about how the woman felt about the incident and what

she wanted to do about it. This video was selected after an extensive internet search for appropriate material. It was deemed to be of the right length, with dialogue that was clear enough to be understood by the majority of English-speaking participants, and a simple scenario. It was also considered to be representative of the kind of incidents experienced by individuals at workplace but mild enough to minimise participant distress.

Components of CI implemented in the AI CI

The interviews were structured according to the UK investigative interview model (PEACE) and Achieving Best Evidence advice (Ministry of Justice 2011). The interview protocols included four ground rules of "explain" phase used in a UK police Tier 1 basic witness interview. These ground rules are; (i) never guess; (ii) report everything; (iii) say if you do not remember; and (iv) say if you do not understand the question.

The first technique applied in the AI CI is the ***context reinstatement***, in which the participants were encouraged to rationally recreate the physical (environmental) and individual (e.g. how they felt at that time) context that existed at the time of the event. Previous research (e.g., Memon and Bull 1991) suggested that any aspect of an environment in which a to-be-remembered event is encoded, in theory, serves as a contextual cue. The second technique applied was to ask the participants to ***report everything*** they can recall regardless of the possibility that they think the details are not important or trivial or even partial or incomplete. For example, *"tell me everything you can, even things you think are not important, and even you cannot remember something completely".*

In the present study, participants were not asked to report from a ***variety of perspectives*** and ***change perspective technique*** (one of the four original CI mnemonic). Do interviewers have to use all parts of the CI? Milne and Bull (2002) analysed the relative adequacy of each of the four

original CI mnemonics in an examination where the participants were adults and children. For all age groups, they found a combination of mental reinstatement of context (MCR) and Report Everything occasioned more recall compared with the individual use of the other techniques. Importantly, there was no significant difference when MCR was used on its own, confirming the determinant role of context reinstatement in the CI. MCR also resulted in proportionally more details when it was followed by an open-ended invitation to elaborate (Hershkowitz et al. 2001). This suggests that interviewers do not always need to use the full procedure to see the benefits of the CI. As such, AI CI uses a combination of MCR and Report Everything techniques.

Walsh and Bull (2012) contend that during an interview rapport is an opportunity to build to establish a working relationship with suspects. In turn, the rapport between the interviewer and suspect can help suspects to supply information more freely. However, the existing literature concerning rapport building, particularly on the context of witness and victim interviews is sparse and provides limited guidance on what constitutes rapport (Vallano and Compo 2015). The interview protocols developed for this study did not include a rapport-building phase as it is typically conceptualised in the Cognitive Interview framework because this was not seen to be possible with a chatbot and our development resources. In an effort to build rapport, we incorporated a greeting within the chatbot and explained the ground rules before moving to the next phase of the interview.

Overall, the purpose of the AI CI was to stick as closely as possible to the original Cognitive Interview protocol within the confines of an AI chatbot that requires no human interviewer. This means that the chatbot imitates the Cognitive Interview, but it is unclear how comparable the experience of using the AI CI is to being subjected to an *in-person* Cognitive Interview.

Procedures

The participants were recruited online, and the entire experiment was conducted online.

Each participant was randomly assigned to one of four conditions: Free Recall (n=32), Questionnaire (n=31), Basic Chatbot (n=32), and AI CI (n=32). Participants were recruited using Mechanical Turk (Buhrmester, Talaifar, and Gosling 2018), and were paid $2 for their time. As soon as participants clicked on the experiment link to participate, they were briefed as to the aims and procedure of the experiment and the purpose of this research. They were also told that they would be presented sequentially with a short video depicting a mild harassment scenario, a reporting system to help them to relate information about the scenario, and a short questionnaire to rate their experience. They were then asked to read and digitally sign the informed consent form, after which they viewed the video. After a short filler task, they were asked to report from the victim's perspective (*as if the harassment had happened to them)* and were randomly assigned to one of the four digital tools (Free Recall, Questionnaire, Basic Chatbot, or AI CI). They were then debriefed. The experiment lasted approximately 20 minutes.

The digital tools were created for reporting details of the harassment scenario the participants had just watched. These tools allowed the participants to type their responses in a text box, and they all had the appearance of a mobile text chat. Thus, each condition had the same appearance (font, background, etc.), but differed regarding the style of the questions. The first condition was a Free Recall tool. It had just one open-ended question "Please describe everything you can remember about what happened. Try not to leave anything out, even if it seems trivial." The participants replied in the text box and clicked ***send*** when they had finished. The second tool was a Questionnaire. In the Questionnaire condition, the free recall question was followed by a series of follow up questions that the participant could see and scroll through on their screen. The

third condition was Basic Chatbot. This condition had the same questions/statements as the Questionnaire condition, but they were presented in a chat format, so that the participant could only see each question once they had answered the preceding one, as in a conversation. The fourth condition was AI CI. The AI CI was identical to the Basic Chatbot condition except that the follow-up questions were not generic, but specific to the information the participant had just provided. For instance, "You referred to *Mike*. Please tell me more about Mike." For an example image of an AI CI follow question and screenshot, see Appendix 3.

Interview coding

Interviews were coded according to a scoring template technique (see Memon et al. 1996). Two researchers reviewed the responses in each interview and coded the responses for details. The interviews themselves were collected using technology making this part of the experiment inherently 'blind'. Each item recalled by participants was scored as either; (i) correct; (ii) erroneous; or (iii) confabulation. The coding was conducted double-blind by one researcher, who was unaware of which condition was being coded. The other researcher also coded the data, and worked from a spreadsheet that did not explicitly identify the condition. Where there were disagreements, the double-blind researcher had the final say.

***Correct responses*:** Two researchers counted the number of items to be remembered and reached a consensus that there were 51 items. The researchers also scored *extra descriptive items* (14) which included those that had been given as contextual information at the start or end of the video in text form (e.g. consequences of the harassment scenario), or items that were both a) not asked for specifically and b) relevant to the incident in the harassment scenario (e.g. person or scene descriptors). These had not been included in the 51 items because they were too numerous.

***Errors and confabulations*:** In this instance, we were interested in the types of mistakes people made when recording their memories of an event, so we considered two types of mistake: errors and confabulations. *Errors* were simple mistakes, such as getting a date or time wrong, or misquoting what someone had said. *Confabulations* were when a participant invented something that had not happened (e.g. there was a witness).

Design

The results were analysed in two ways. First, using orthodox statistics (one-way ANOVAs) to examine the main effects. Then, Bayesian hypothesis testing was conducted to evaluate the theories in terms of strength of evidence. Bayesian hypothesis testing was considered to be appropriate to evaluate the strength of evidence for the alternative hypothesis over the null, as differences between conditions were subtle. Because the participants had a very limited exposure to the scenario (the video only lasted one minute and 45 seconds). This short stimulus video was not able to elicit differences in the number of recalled items. The Bayes factors thus allowed more nuanced inferences to be made about the data that a) did not depend on power calculations and b) allowed to test support for the null hypothesis. Bayes factors can be used to test whether the data (i) support the null hypothesis (H0), (ii) strengthen support for the alternative hypothesis (H1), or (iii) whether there is no evidence either way. They also challenge perceptions of the importance of power, as they indicate that a high-powered non-significant result is not always evidenced to support the H0, but a low-powered non-significant result might be. Similarly, a significant high-powered result might not be substantial evidence of H1 (Dienes and McLatchie 2017).

To calculate a Bayes factor, one needs a model of H0 (usually that there will be no difference between means), a specified model of H1 (usually from the mean difference in a previous study), and a model of the data. This means that the Bayes factor provides a continuous

measure of evidence strength for H1 over H0, rather than a sharp boundary of significance (Dienes and McLatchie, 2017). Previous research findings into "cognitive" versus "standard" interviews was used to specify the hypothesis. Cognitive interviews were found to elicit a median of 34% more information than standard interviews (Koehnken et al. 1999). Therefore, the *SD* was set to *x* = 34% of the highest score in the present experiment. This figure was calculated separately for each set of comparisons (according to the highest score for that set).

Results:

Two one-way ANOVAs were performed to see whether the number of responses was different when participants used different reporting tools. There was one between-subjects factor: report type (with four levels: Free recall (FR), Questionnaire, Basic Chatbot, and AI CI). The dependent variables were the number of correct, erroneous, or confabulated responses. Bayes factors (B) were also used to determine how strong the evidence for the alternative hypothesis (H1) was over the null (H0) (Singh n.d.). For additional results related to all Baysian analyses, please refer to Appendix 2.

Correct Responses

For correct responses, we used the *SD* (34%) to test this when conducting Bayesian hypothesis testing (as Cognitive interviews were found to elicit a median of 34% more information than a standard interview). For the first analyses (overall correct responses), the *SD* was set to 6.08.

Inspection of figure 2 revealed that for ***correct responses*** overall, there was an effect of each type reporting tool. This is supported by a one-way ANOVA, F (3, 126) = 2.69, $p < .05$, r = .26, (Free Recall, M = 14.17, SE = 0.93; Questionnaire, M = 16.68, SE = 0.81; Basic Chatbot, M = 15.83, SE = 0.84; AI CI, M = 17.88, SE = 1.14). Bonferroni post hoc analyses revealed that AI CI elicited more correct responses than Free Recall, p = .04, B_H = 182.55. There were no considerable

effects between Basic Chatbot and Free Recall, $p = 1$, $B_H = 1.54$; between Questionnaire and Free Recall, $p = .37$, $B_H = 16.65$; between Questionnaire and Basic Chatbot, $p = 1$, $B_H = 0.07$; between Questionnaire and AI CI, $p = 1$, $B_H = 0.59$; and between Basic Chatbot and AI CI, $p = .78$, $B_H = 2.54$).

The combined analyses thus indicated that there was evidence that both Questionnaire CI and AI CI elicited more correct items overall than Free Recall tool. The results also indicated that there was evidence to support the null when it came to comparisons between Questionnaire CI and Basic Chatbot CI.

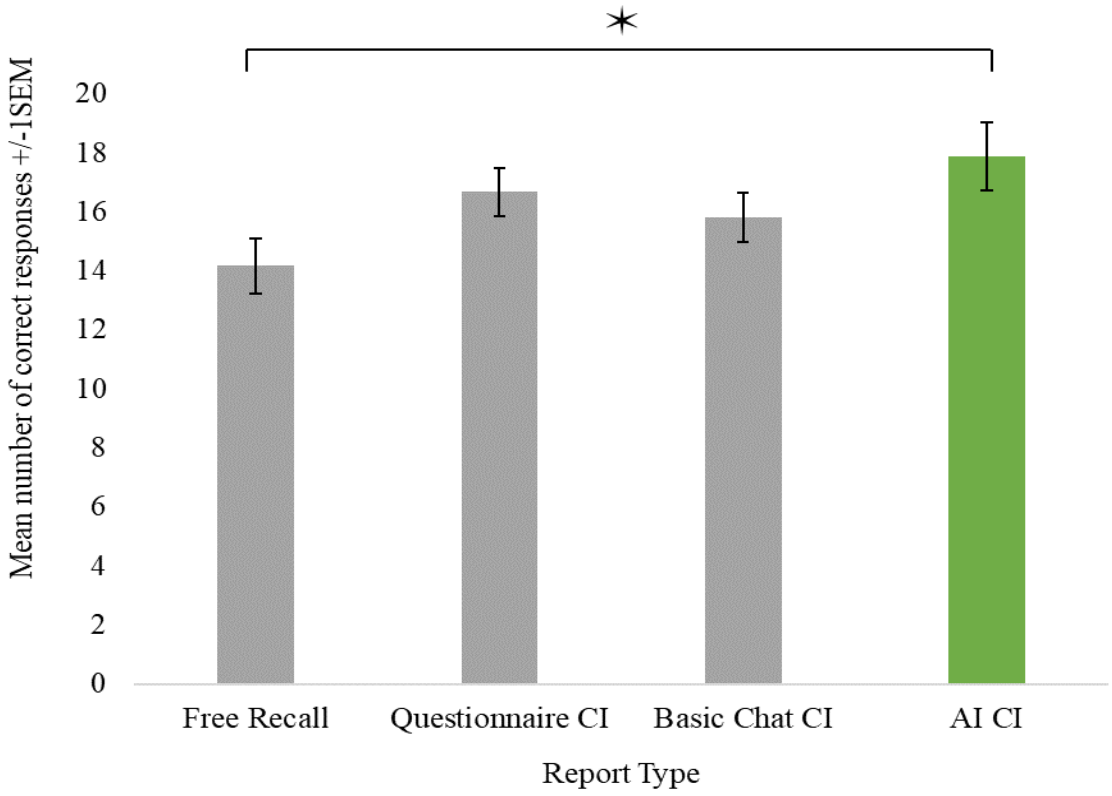


Figure 2. Number of *correct* items recalled as a function of reporting tool.

***Descriptive items*:** These were items that were either given as contextual information at the start or end of the video in text form, or items that were both a) not asked for specifically and b)

relevant to the incident in the harassment scenario. When using Bayesian hypothesis testing, the focus was only on comparisons between the AI CI and the other conditions (as we predicted that NLP would help participants to recall extra descriptive items), and the *SD* was set to 1.46 (34% of the highest score – as Cognitive interviews were found to elicit a median of 34% more information than standard interviews; Koehnken et al. 1999) for these comparisons.

Figure 3 shows a significant effect of condition on the number of descriptive items recalled, F (3,126) = 11.22, $p < .001$, $r = .48$, (Free Recall, $M = 1.00$, $SE = 0.35$; Questionnaire, $M = 1.91$, $SE = 0.22$; Basic Chatbot, $M = 1.97$, $SE = 0.33$; AI CI, $M = 4.09$, $SE = 0.57$). Bonferroni post hoc analyses revealed that AI CI elicited more descriptive items than Free Recall ($B_H = 50214.61$), Questionnaire CI ($B_H = 3059.80$), and Basic Chatbot ($B_H = 57.14$), all ps < .001. There were no other effects (between Basic Chatbot and Free Recall, $p = .57$; between Questionnaire and Free Recall, $p = 1$; and between Questionnaire and Basic Chatbot, $p = 1$). AI CI performed better than the other conditions in eliciting descriptive details. The Bayes factor supported the results for these comparisons.

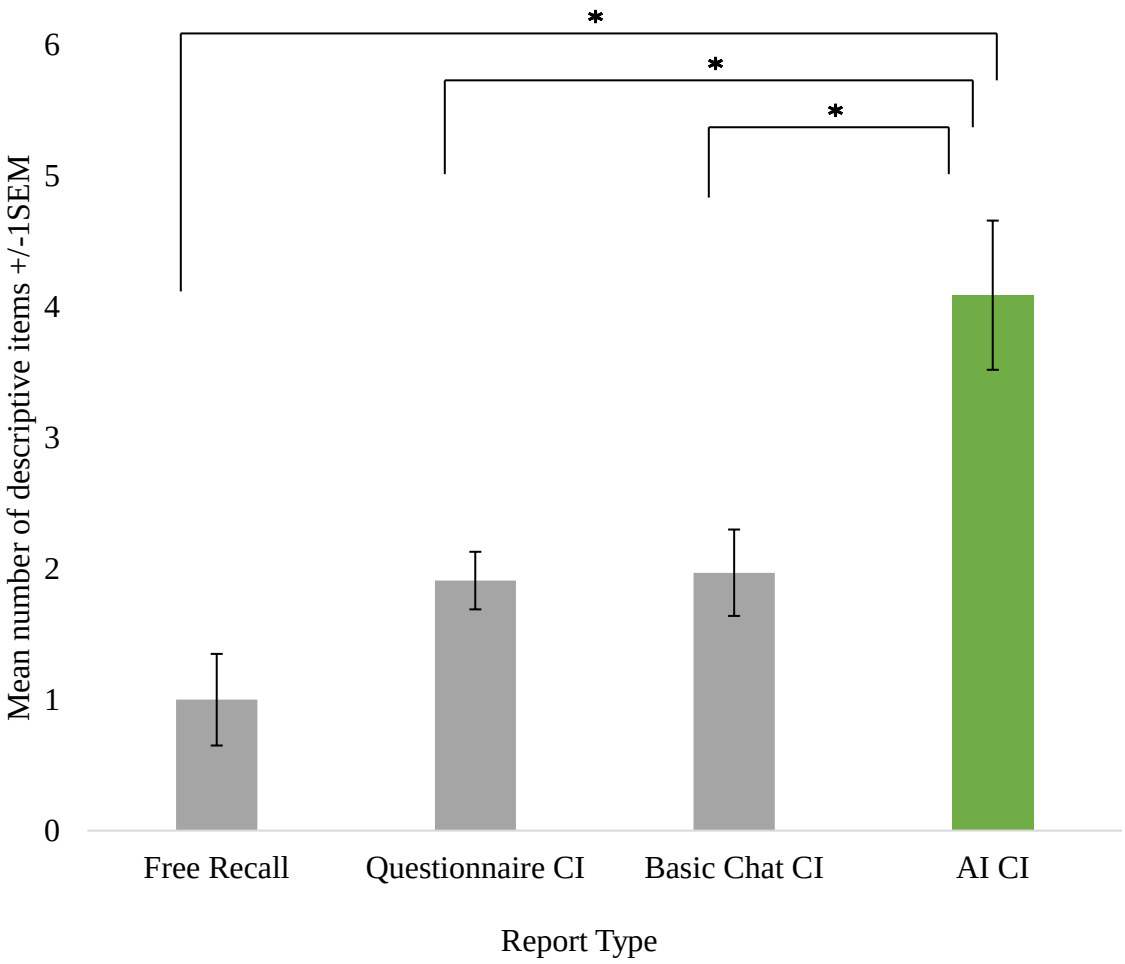


Figure 3. Number of *descriptive* items recalled correctly as a function of reporting tool.

Incorrect Responses

For ***incorrect responses***, we expected the number of mistakes to be *fewer* as the sophistication of the reporting tool improved, so when conducting Bayesian hypothesis testing, the *SD* was set to $x = 0.83$.

Inspection of figure 4 reveals that there was a significant effect of reporting tool, F (3, 126) = 3.47, p = .02, r = .16, (Free Recall, M = 1.23 SE = 0.30; Questionnaire, M = 2.44, SE = 0.33; Basic Chatbot, M = 1.90, SE = 0.31; AI CI, M = 1.48, SE = 0.20). Bonferroni post hoc analyses

revealed that Questionnaire CI elicited more incorrect responses than Free Recall, $p = .02$, $B_H = 51.17$, although the differences were small in absolute terms. There were no other significant effects (between Basic Chatbot and Free Recall, $p = .67$, $B_H = 1.54$; between AI CI and Free Recall, $p = 1$, $B_H = 0.29$; between Questionnaire CI and Basic Chatbot, $p = 1$, $B_H = 11.99$; between Questionnaire and AI CI, $p = .10$, $B_H = 296.03$; and between Basic Chatbot and AI CI, $p = 1$, $B_H = 1.99$).

The participants in the Questionnaire CI condition elicited more incorrect responses overall than Free Recall. However, the Bayes factors indicated substantial evidence that Questionnaire CI also encouraged more incorrect responses than both AI CI and Basic Chatbot CI. Bayes factors allowed us to conclude that there was no difference in the number of incorrect responses between AICI and Free Recall, indicating that these two conditions encouraged accuracy more than the other two. Finally, for incorrect responses, Bayes factors indicated that there were *no* differences between the Basic Chat CI and Free Recall, or between the Basic Chat CI and AI CI, as the results were insensitive.

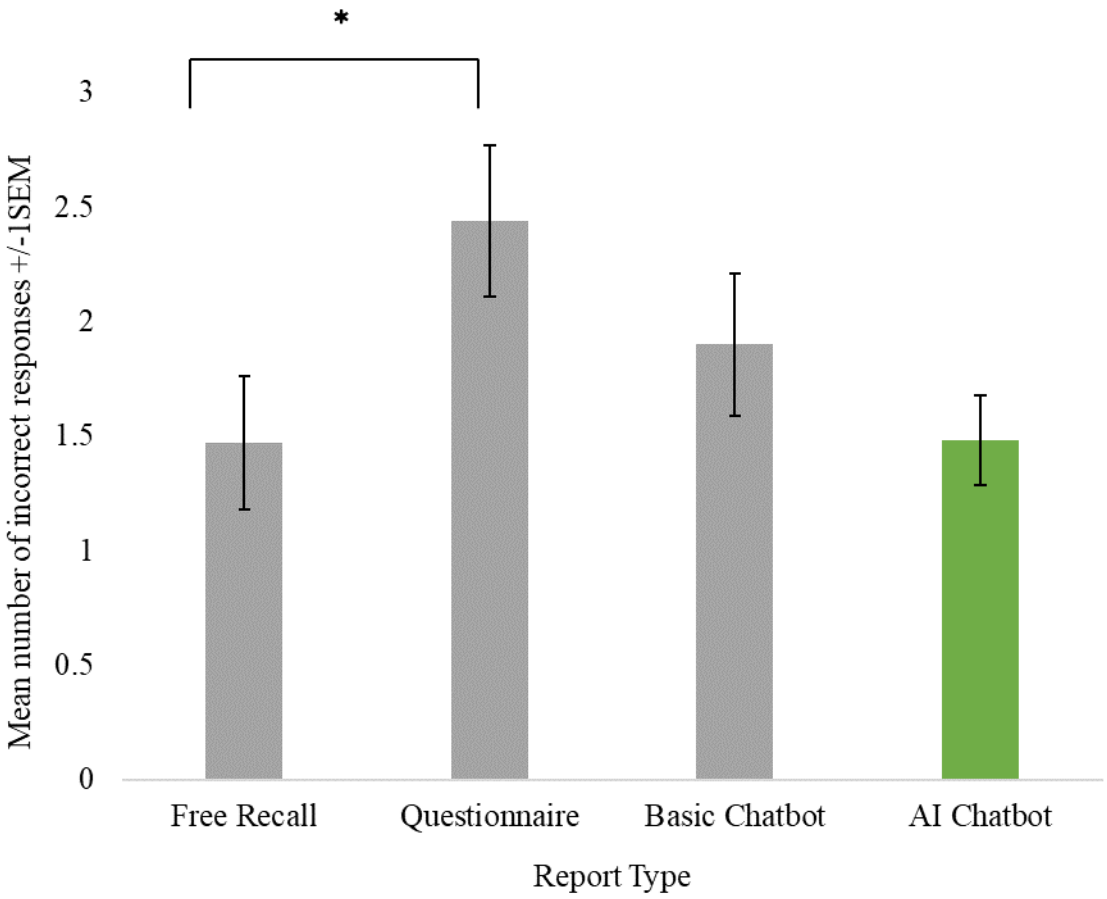


Figure 4. Number of *incorrect* items recalled as a function of reporting tool.

Finally, an exploration was conducted examining whether people recalled items incorrectly for different reasons when accounting for the different conditions. For this, the incorrect responses were separated into *errors* and *confabulations*.

***Errors**:* For these analyses, again, the focus was on comparisons between the AI CI and the other conditions when conducting Bayesian hypothesis testing, and the *SD* was set to 0.28 for these comparisons.

Inspection of figure 5 shows that there was no effect of condition on number of errors; $F(3,126) = 1.17$, $p = .33$, (Free Recall, $M = 0.60$, $SE = 0.18$; Questionnaire, $M = 0.76$, $SE = 0.14$; Basic Chatbot CI, $M = 0.83$, $SE = 0.19$; AI CI, $M = 0.45$, $SE = 0.12$). However, the Bayes factor

indicated that the results between Basic Chatbot CI and AI CI, and Questionnaire CI and AI CI supported the alternative hypothesis (B_H = 11.30 and B_H = 8.61 respectively), and those between Free Recall and AI CI supported the null hypothesis, B_H = 0.26.

Therefore, while significance testing suggested that it made no difference which reporting tool participants used, Bayesian analysis indicated that participants using AI CI made fewer errors than those using Questionnaire CI or Basic Chatbot CI.

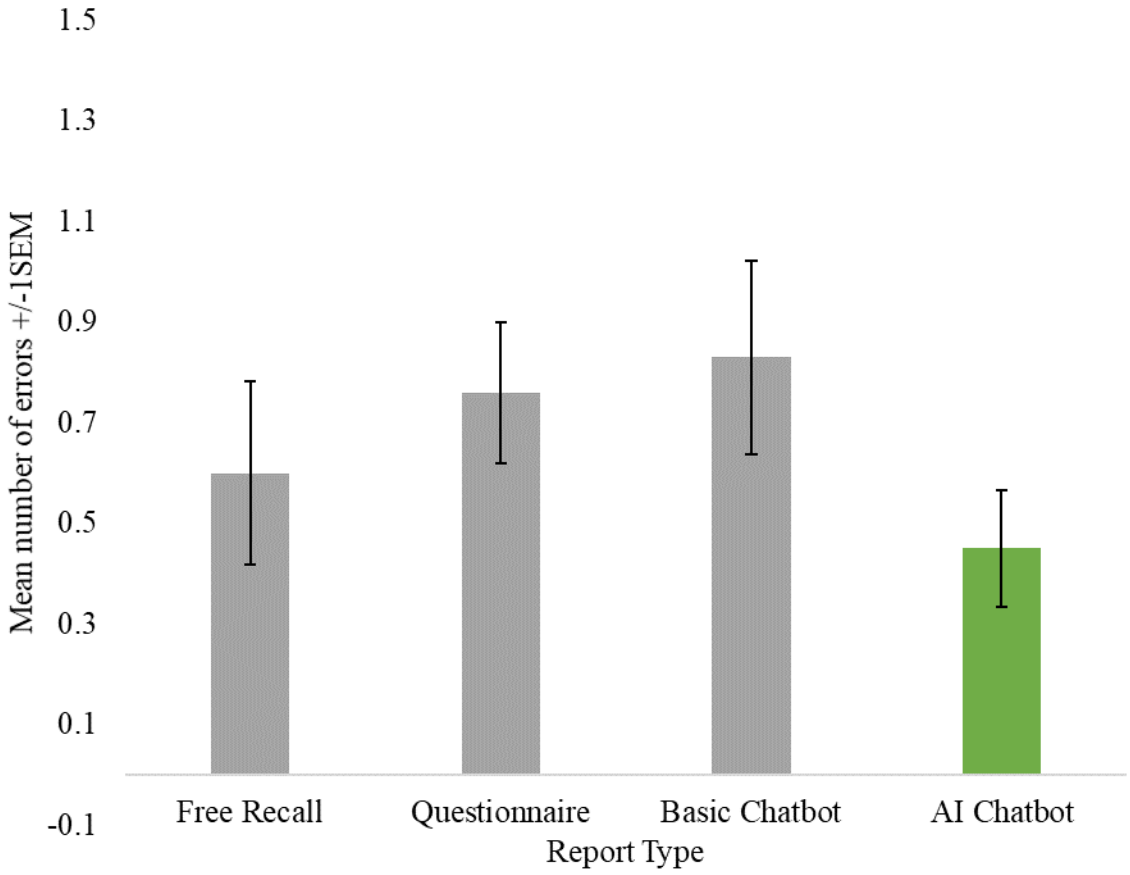


Figure 5. Number of *errors* recalled incorrectly as a function of reporting tool.

Confabulations*:* For the final analyses, we also focused on comparisons between the AI CI and the other conditions for the Bayesian hypothesis testing, and the *SD* was set to 0.57 for these comparisons.

As presented in figure 6, there was a significant effect of condition on the number of confabulations, F (93,126) = 4.52, p = .01, r = 32, (Free Recall, M = 0.63, SE = 0.17; Questionnaire, M = 1.68, SE = 0.28; Basic Chatbot CI, M = 1.07, SE = 0.20; AI CI, M = 0.97, SE = 0.14). Bonferroni post hoc analyses revealed that Questionnaire CI elicited significantly more confabulations than Free Recall, p = .01. However, the absolute difference was small. There were no other significant differences (between Basic Chatbot CI and Free Recall, p = .92; between AI CI and Free Recall, p = 1; between Questionnaire CI and Basic Chatbot CI, p = .23; between Questionnaire CI and AI CI, p = .09; and between Basic Chatbot and AI CI, p = 1).

However, the Bayes factor indicated that the results between Free Recall and AI CI, and between Questionnaire CI and AI CI, B_H = 4.30, and B_H = 106.08 supported the alternative hypothesis respectively. The comparison between Basic Chatbot CI and AI CI, B_H = 0.48 was insensitive.

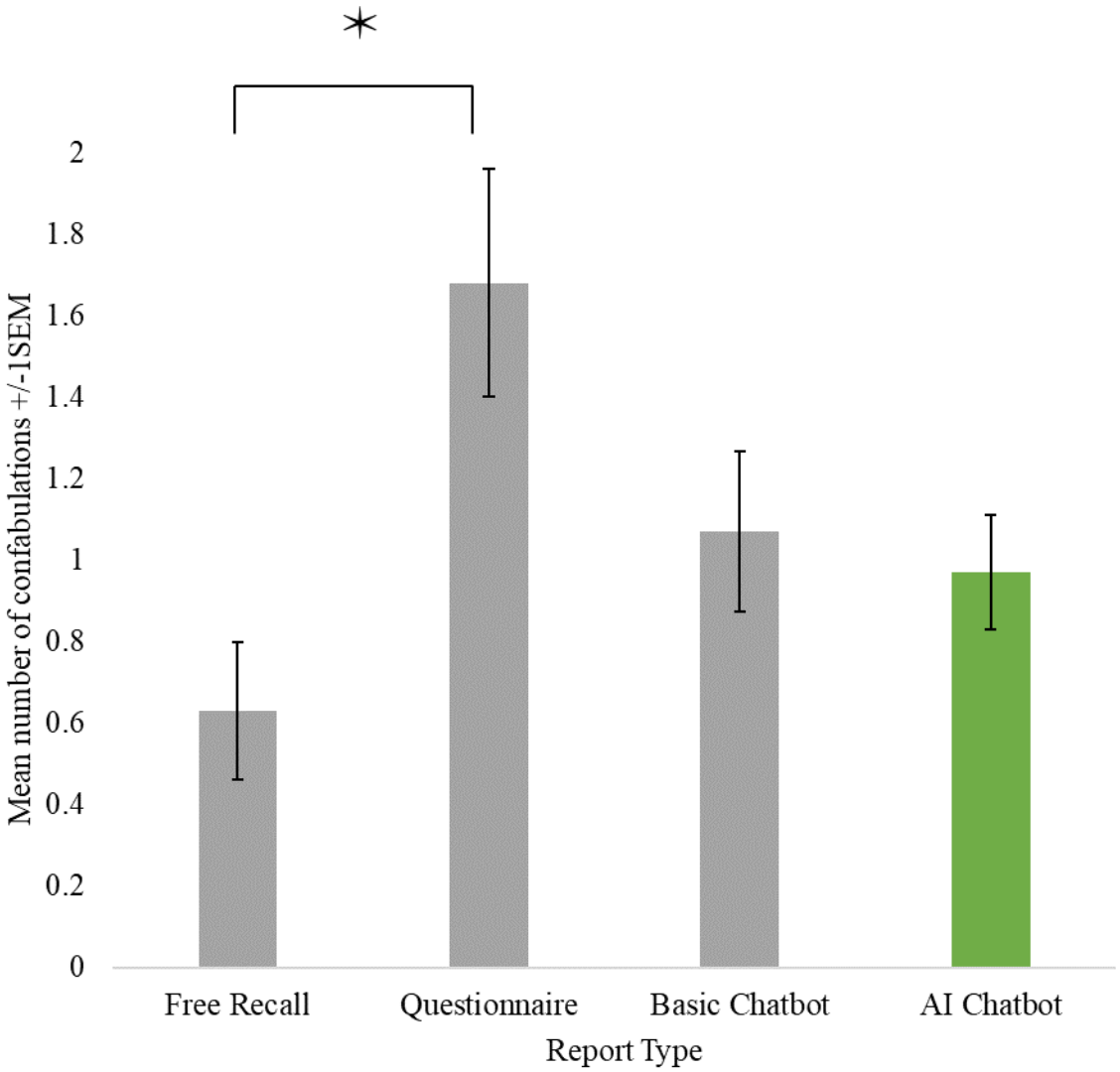


Figure 6. Number of *confabulations* recalled incorrectly as a function of reporting tool.

Discussion

The present study examined whether AI CI would help to elicit accurate recall when participants recorded a harassment scenario, compared with other online reporting mechanisms. First considering *correct* responses, the results showed that the AI CI helped people record more correct information overall than other conditions. Despite the limited exposure to the scenario (the video only lasted one minute and 45 seconds), AI CI participants recorded an average of four extra

correct responses compared with other conditions. Bayesian hypothesis testing was considered to be appropriate to evaluate the strength of evidence for the alternative hypothesis over the null, as differences between conditions were subtle. The stimulus was a short video that was not able to elicit dramatic differences in the number of recalled items. Bayes factors allowed us to make conclusions that were not possible using orthodox statistics, and in some instances, supported the statistics, adding weight to the implications. For instance, significance testing indicated that overall, the AI CI helped people to recall more items overall, that the AI CI was better than the other conditions at eliciting description, and the Bayes factors supported these results.

The accuracy of witnesses' recall increases if they communicate only those recollections they are certain of and refrain from guessing and judging (Koriat and Goldsmith 1996). Research has identified an investigator bias effect, where interviewers are biased towards thinking that an interviewee is deceitful. Prior experience and training are suggestively correlated with a tendency to judge interviewees as deceitful rather than truthful, although experience and training are not correlated with the accuracy of interviewers' judgments (Meissner and Kassin 2002). A bias of this kind could lead to an 'accusatorial' style of forensic interviewing (Mortimer and Shepherd 1999) where interviewers use a 'confirmatory' strategy to affirm their beliefs (Minhas et al. 2017). Such interviewing methods can also increase the likelihood of false confessions (Gudjonsson 2018; Narchet et al. 2011). As such, the AI interviewing tool (which mimics the cognitive interview) provide an unbiased, and non-judgmental facility to obtain accurate accounts and help people to record more correct information.

Similarly, the question format also influences recall accuracy. Responses to open-ended questions are more accurate than closed and leading questions. An over-riding principle of the CI then is to conduct the interview primarily by asking open-ended questions, and appropriate

open/closed follow up questions (Fisher and Geiselman 2010). One of the possible explanations, why AI CI elicited more accurate responses, could be because it follows a structured, scripted interviewing protocol with consistent open questions followed by appropriate follow-up questions. Research shows the interviews conducted with less structured protocols are more susceptible to suggestibility and bias effects (see Santtila et al. 2004). As such, AI CI may also potentially reduces the suggestibility and bias effects.

As the types of items that could be recalled were different, we then divided them into four different memory types, so that we could see what type of information was recalled most successfully in each condition. When we examined *descriptive* items, it was found that people in the AI CI condition performed better than those in the other conditions, recording about twice as many items. Research shows that memory is fallible, and people are often unable to recall sufficient forensic details such as person descriptors (Kebbell and Milne 1998), so this finding could be seen as promising. The CI is a witness-centred approach with a transfer of control to the interviewee (Fisher and Gielseman 2010). As such, the communication components of the AI CI will heighten the witness's sense of control (by mimicking the CI in a humane way without a real human) in turn, restoring some of the power that was lost in the victimisation which might result in recalling more descriptive items. This is in line with previous research (e.g., Gabbert et al. 2009; MacLean et al. 2019) which found that Self-administered interviews help to recall more correct information as compared to traditional interviews.

A self-administered interviewing tool is only useful if it does not increase the number of incorrect responses. In the present study, Bayesian hypothesis testing revealed that the AI CI elicited the fewest incorrect responses, which was largely driven by people in the Questionnaire CI condition being more likely to *confabulate*. As such, findings from the current research shows

that the AI CI helped people to record more information than the other digital tools examined during this study without compromising on accuracy, and was good at encouraging description. Therefore, the present study's findings are promising. Overall, findings indicate that an AI-based chatbot (such as AI CI) could be an effective means of gathering high-quality digital evidence immediately after an incident.

Employment Opportunity Commission (EEOC 2016) found that approximately 70 per cent of victims did not report to their workplaces even though staying silent has health and career consequences (Cortina and Magley 2003) and isolates the victim (Herbenick et al. 2019). Research studies found that employees do not report workplace harassment due to: (i) fear of retaliation or other negative consequences; (ii) fear of not being believed; (iii) worry about being judged; (iv) feel it's embarrassing to report; and (v) know there is a conflict of interest. A report from the British *Equality and Human Rights Commission* (2018) indicates that anonymous online reporting tools can improve reporting processes and reduce reporting barriers. However, anonymity and good policies are arguably insufficient if the option is poorly designed. The reporting process itself also needs to create a strong case, eliciting high-quality evidence that would be taken seriously and acted upon. AI CI has tremendous value in making evidence-based memory interviews for important emotional events accessible to a broad audience online 24/7. It is also an immediate and practical solution to an issue of increasing global concern – the underreporting of workplace harassment and discrimination.

Limitations

There are also limitations of using an AI CI, most notably the added difficulty of establishing rapport with the individual (which is also a step in the full cognitive interview). Additionally, there

is the inability of an AI to express empathy or emotional support which some individuals may want or need, and the inability of the AI to identify critical situations and provide help where an individual may need an immediate resource. There can also be the potential for awkwardness, because it may ask follow-up questions that are not in line with an ongoing conversation, potentially having misidentified keywords. While the purpose of the present study was to examine the viability of a Cognitive Interview chatbot in comparison to other digital tools it would also be useful for future research to examine how the information collected by the AI CI compares to a human interviewer trained in the Cognitive Interview. Future research might also benefit from larger sample sizes.

Conclusions

AI CI allows witnesses and victims to take a more active role in the interview. This may not only increase the amount of information retrieved but also gives the witness and victim a voice in the investigative process, which potentially promotes a sense of self-efficacy and control over the interview process (Fisher and Geiselman 2010). The AI CI could be seen as a useful Cognitive Interview tool to effectively and efficiently record high-quality evidential statements from victims and witnesses. This is particularly valuable because of the scalability and access that AI can offer. The AI CI potentially eliminates inconsistent interviewer effects, provides an un-biased, confidential, and safe place to record statements immediately after an incident.

Ethics statement: The present study was approved by the author's home university and run in accordance with the British Psychological Society code of ethical conduct. A potential conflict of interest has been declared throughout the ethics process because this study was funded by a San Francisco based company called All Turtles on behalf of Spot, and one of the three authors of this

paper is the co-creator of Spot. Spot is an AI chatbot that was based in part on the results of the present research but has since been modified for broader purposes. The most recent version of Spot can be accessed for free by individuals via https://app.talktospot.com/ . The AI CI used in this study was specifically designed for research purposes, and if you would like to conduct research using this version it is recommended that you contact one of the authors of the present paper.

References

Brandtzaeg B, Følstad A (2017 November). Why people use chatbots. In *International Conference on Internet Science* (pp. 377-392). Springer, Cham. https://doi.org/10.1007/978-3-319-70284-1_30

Buhrmester M. D, Talaifar S, Gosling, D (2018). An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. *Perspectives on Psychological Science*, *13*, 149-154. https://doi.org/10.1177/1745691617706516

Bull R, (2013). What is 'believed' or actually 'known' about characteristics that may contribute to being a good/effective interviewer? *Investigative Interviewing: Research and Practice (II---RP),* 5, 128-143

Carter M, Thompson N, Crampton P, Morrow G, Burford B, Gray C, Illing J, (2015). Workplace bullying in the NHS: Prevalence, impact and barriers to reporting. *BritishPsychological Society North East Branch Bulletin*, *2015*, 30-35.

Cortina M, and Magley J, (2003) Raising voice, risking retaliation: Events following interpersonal mistreatment in the workplace. *Journal of occupational health psychology*, *8*(4), p.247.

Dando C, Oxburgh G, (2016). Empathy in the field: Towards a taxonomy of empathic communication in information gathering interviews with suspected sex offenders. The European Journal of Psychology Applied to Legal Context, 8, 27-33.

Dienes Z, Mclatchie N, (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic bulletin & review*, *25*, 207-218. DOI 10.3758/s13423-017-1266z

Fisher R, Geiselman R, (2010). The cognitive interview method of conducting police interviews: Eliciting extensive information and promoting therapeutic jurisprudence. *International journal of law and psychiatry*, *33*(5-6), 321-328. https://doi.org/10.1016/j.ijlp.2010.09.004

EEOC, (2016). Select task force on the study of harassment in the workplace. www.eeoc.gov/eeoc/task_force/harassment/report.cfm. Accessed January 16, 2018

EHRC (2018), retrieved 4th May, from https://www.equalityhumanrights.com/en/publication-download/turning-tables-ending-sexual-harassment-work

Fisher R, Milne R, Bull R, (2011). Interviewing cooperative witnesses. *Current Directions in Psychological Science*, *20*, 16-19. https://doi.org/10.1177%2F0963721410396826

Gabbert F, Hope L, Fisher R, (2009). Protecting eyewitness evidence: Examining the efficacy of a self-administered interview tool. *Law and human behavior*, *33*, 298 307. DOI 10.1007/s10979-008-9146-8

Gabbert F, Memon A, Allan K, (2003). Memory conformity: Can eyewitnesses influence each other's memories for an event?. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *17*, 533-543. https://doi.org/10.1002/acp.885

Gudjonsson G, (2018). *The psychology of interrogations and confessions: A handbook*. London: John Wiley & Sons.

Herbenick D, van Anders M, Brotto A, Chivers L, Jawed-Wessel S, Galarza J, (2019). Sexual harassment in the field of sexuality research. Archives of sexual behavior, 48(4), 997-1006.

Kaplan R, Van Damme I, Levine L, Loftus E, (2016). Emotion and false memory. *Emotion Review, 8*, 8-13. https://doi.org/10.1177%2F1754073915601228

Kebbell M, Milne R, (1998). Police officers' perceptions of eyewitness performance in forensic investigations. *The journal of social psychology*, *138*, 323-330. https://doi.org/10.1080/00224549809600384

Köhnken G, Milne R, Memon A, Bull R, (1999). The cognitive interview: A metaanalysis. *Psychology, Crime and Law*, *5*(1-2), 3-27. https://doi.org/10.1080/10683169908414991

Lamb M, Hershkowitz I, Orbach Y, Esplin P, (2008). Factors affecting the capacities and limitations of young witnesses. In *Tell Me What Happened: Structured Investigative Interviews of Child Victims and Witnesses* (pp. 19-61). Chichester, UK: Wiley.

London K, Henry L, Conradt T, Corser R, (2013). Suggestibility and individual differences in typically developing and intellectually disabled children. *Suggestibility in legal contexts: Psychological research and forensic implications*, 129-148.

Meissner C, Kassin S, (2002). "He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior*, *26*, 469-480. Doi:10.1023/a:1020278620751

Memon A, Holley A, Wark L, Bull R, Koehnken G, (1996). Reducing suggestibility in child witness interviews. *Applied Cognitive Psychology, 10,* 503-518.

Memon A, Meissner C, Fraser J, (2010). The Cognitive Interview: A meta-analytic review and study space analysis of the past 25 years. *Psychology, public policy, and law*, *16*,340-372.

Milne R, Bull R, (1999) *Investigative interviewing: Psychology and Practice.* Chichester: John Wiley & Sons Ltd.

Milne R, Bull, R, (2016). Investigative interviewing: investigation and probative value. *The Journal of Forensic Practice, 18*. DOI: 10.1108/JFP-01-2016-0006

Minhas R, Walsh D, Bull R, (2017). Developing a scale to measure the presence of possible prejudicial stereotyping in police interviews with suspects: The Minhas Investigative Interviewing Prejudicial Stereotyping Scale (MIIPSS). *Police Practice and Research, 18*, 132-145. https://doi.org/10.1080/15614263.2016.1249870

Ministry of Justice, (2011). Achieving best evidence in criminal proceedings: Guidance on interviewing victims and witnesses, and guidance on using special measures. London: Ministry of Justice.

Mortimer A, Shepherd E, (1999). Frames of mind: Schemata guiding cognition and conduct in the interviewing of suspected offenders. *A. Memon & R. Bull (Eds.), Handbook of the psychology of interviewing*, 293-315.

Murphy G, Greene C, (2016). Perceptual load affects eyewitness accuracy and susceptibility to leading questions. *Frontiers in psychology*, *7*, 1322. https://doi.org/10.3389/fpsyg.2016.01322

Narchet F, Meissner C, Russano M, (2011). Modeling the influence of investigator bias on the elicitation of true and false confessions. *Law and human behaviour, 35,* 452-465. DOI: 10.1007/s10979-010-9257-x

Perfect T, Wagstaff G, Moore D, Andrews B, Cleveland V, Newcombe S, Brown L, (2008). How can we help witnesses to remember more? It's an (eyes) open and shut case. *Law and Human Behavior, 32,* 314-324. Doi: 10.1007/s10979-007-9109-5

Poole D, Lamb M, (1998). Investigative interviews of children: A guide for helping professionals. Washington, DC, US: American Psychological Association.

Prendinger, H., & Ishizuka, M. (Eds.). (2013). *Life-like characters: tools, affective functions, and applications*. Springer Science & Business Media.

Rahman A., Al Mamun A, Islam A, (2017). Programming challenges of chatbot: Current and future prospective. In *Humanitarian Technology Conference (R10-HTC), 2017 IEEE Region 10* (pp. 75-78). IEEE. https://doi.org/10.1109/R10-HTC.2017.8288910

Read J, Connolly D, (2017). The effects of delay on long-term memory for witnessed events. In M. P. Toglia, J. D. Read, D. F. Ross and R. C. L. Lindsay (Eds.), *Handbook of eyewitness psychology: Volume 1: Memory for events* (pp. 117 155). Mahway, NJ: Lawrence Erlbaum Associates Inc.

Ridley A, (2013). Suggestibility: A History and Introduction. In A. M. Ridley, F. Gabbert and D. J. La Rooy (Eds.), Suggestibility in Legal Contexts: Psychological Research and Forensic Implications (pp. 1-19). England: Wiley-Blackwell.

Rossmo, D. K. (2016). Case rethinking: a protocol for reviewing criminal investigations. *Police Practice and Research, 17,* 212-228. https://doi.org/10.1080/15614263.2014.978320

Santtila, P., Korkman, J., & Sandnabba, N. K. (2004). Effects of interview phase, repeated interviewing, presence of a support person, and anatomically detailed dolls on child sexual abuse interviews. *Psychology, Crime & Law, 10,* 21-35. Doi:10.1080/1068316021000044365

Shaw J, Porter S, (2015). Constructing rich false memories of committing crime. *Psychological science, 26,* 291-301. https://doi.org/10.1177%2F0956797614562862

Shawar B, Atwell E, (2007). Chatbots: are they really useful? In *Ldv Forum* (Vol. 22, No.1, pp. 29-49).

Shawar B, Atwell E, (2005). Using corpora in machine-learning chatbot systems. *International journal of corpus linguistics, 10,* 489-516. https://doi.org/10.1075/ijcl.10.4.06sha

Singh A, (n.d.), Bayes Factor (Dienes) calculator, retrieved 5th September, 2018, from https://medstats.github.io/bayesfactor.html.

Stein L, Memon A, (2006). Testing the efficacy of the cognitive interview in a developing country. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 20,* 597-605. https://doi.org/10.1002/acp.1211

Taylor D, Dando C, (2018) Eyewitness Memory in Face-to-Face and Immersive Avatar-to Avatar Contexts. Front. Psychol. 9:507. https://doi.org/10.3389/fpsyg.2018.00507

Tuckey M, Brewer N, (2003). The influence of schemas, stimulus ambiguity, and interview schedule on eyewitness memory over time. *Journal of Experimental Psychology: Applied, 9,* 101-118.

Turtle J, Yuille J, (1994). Lost but not forgotten details: Repeated eyewitness recall leads to reminiscence but not hypermnesia. *Journal of Applied Psychology, 79,* 260.

Vallano J, Compo N, (2011). A comfortable witness is a good witness: Rapport building and susceptibility to misinformation in an investigative mock-crime interview. *Applied Cognitive Psychology*, *25*, 960-970. https://doi.org/10.1002/acp.1789

Vallano J,Schreiber Compo N, (2015). Rapport-building with cooperative witnesses and criminal suspects: A theoretical and empirical review. *Psychology, Public Policy, and Law*, *21*, 85-89.

Walsh D, Bull R, (2012). Examining rapport in investigative interviews with suspects: Does its building and maintenance work? *Journal of police and criminal psychology*, *27*(1), 73 84. DOI: 10.1007/s11896-011-9087-x

Westera N, Kebbell M, Milne B, (2011). Interviewing witnesses: do investigative and evidential requirements concur? *The British Journal of Forensic Practice*, *13*, 103-113. https://doi.org/10.1108/14636641111134341

Wixted J, Ebbesen E, (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & cognition*, *25*, 731-739. DOI: 10.3758/BF03211316

YouTube, n.d., retrieved, 3rd April 2018, from https://www.youtube.com/watch?v=kg7k5x--k8o&t=22s.

Appendices

Appendix 1

Description of AI NLP Training

Training the AI. In order to help the AI learn which words are important and in what contexts, tables are created manually that we feed into the AI. We provide examples and it figures out the

relationship between words. The more we have the more the links start to be concrete. For example, we manually indicate to the AI that “boss” is a “job role”, so that it learns to ask follow-up questions about “boss”. For some categories of words, training models already exist. For example, we use a standard library of names. But no standard library exists for words related to workplace harassment and discrimination. Because of this, we have created three particular training datasets of words and phrases to train our AI.

Group 1 relates times and dates. For this we have manually filtered a pre-existing database, filtering out words that were too general for our context like “a few” and other broad numerical descriptions that were not appropriate. *Group 2* relates to locations. Here we have created a completely custom library based on workplace-related terms, like “office” or “boardroom”. *Group 3* relates to people – including roles, job titles, and names. We have created a bespoke library of workplace related descriptions like “she is my boss” or “colleague”. The names library is a standard database that has been applied unmodified.

Our own training dataset of about 1000 sentences was created using four main stages. The first stage consisted of brainstorming what we expected to be asked, resulting in about 100 sentences. In the second stage, we harvested words and phrases from news articles described accounts of workplace harassment and discrimination. This provided different syntax and word choice and added to our database. In stage three, we used about 200 reports submitted to the team explicitly for research purposes (from talktospot.com) to add to our database (Note that although many reports have been created using talktospot.com, we do not have access to them unless they are explicitly sent to the research team. This means we cannot assess the quality of the AI in these interactions).

Currently, in stage four, we are developing industry-specific words and phrases based on the industries that are using our tool. Ultimately, the database will be continuously evolving, and the AI should become increasingly attuned to the relevant words and their contexts to improve the follow-up questions and the user experience.

Appendix 2

Analyses using the Bayes Factor.

Introduction

Bayes factors are useful for assessing the strength of evidence of a theory, and allow researchers to draw different conclusions from those that can be inferred from orthodox statistical methods alone. Orthodox statistics model the null hypothesis (H_0), generally testing if there is no difference between means. They reveal whether there is a statistical difference between means, but nothing else. Bayes factors can be used to make a three-way distinction, by testing whether the data either support the null hypothesis (H_0), whether they strengthen support for the alternative hypothesis (H_1), or whether there is no evidence either way. Bayes factors also challenge perceptions of the importance of power that are used in statistics, as they indicate that a high-powered non-significant result is not always evidence to support the H_0, but a low-powered non-significant result might be. Similarly, a high-powered significant result might not be substantial evidence of H_1. Finally, using Bayes one can specify the hypothesis in a way that is not possible with a *p* value (Dienes & McLatchie, 2017).

To calculate a Bayes factor, one needs a model of H_0 (usually that there will be no difference between means), a model of H_1 (which needs to be specified, usually from the mean difference in a previous study) and a model of the data. This means that the Bayes Factor provides a continuous measure of evidence strength for H_1 over H_0, rather than a sharp boundary of

significance. However, as a Bayes factor of 3 often aligns with a $p < .05$, a Bayes factor of 3 or more is usually understood as substantial evidence in support of H1. For symmetry, substantial support for H_0 is usually understood as a Bayes Factor of < 1/3 (Dienes & McLatchie, 2017).

Therefore, in the present research, as well as examining the main effects with statistics, we evaluated the theories in terms of strength of evidence, using Bayesian hypothesis testing. Bayes factors seemed appropriate as the difference between the conditions was designed to be subtle, and the video stimuli were short, so we were expecting non-significant results in some comparisons. The Bayes Factors also allowed us to make more nuanced inferences about the data that did not depend on power calculations.

Methods

For our analyses, Bayes factors (*B*) were used to determine how strong the evidence for the alternative hypothesis was over the null (Singh, n.d.). B_H (0, *x*), indicates that predictions of H1 were modelled as a half-normal distribution with a standard deviation (*SD*) of *x* (Dienes & McLatchie, 2018). We used previous research into "cognitive" versus "standard" interviews to specify our hypothesis. This showed that cognitive interviews were found to elicit a median of 34% more information than standard interviews (Koehnken, Milne, Memon, & Bull, 1999). Therefore, the *SD* was set to *x* = 34% of the highest score in the present experiment. This figure was calculated separately for each set of comparisons (according to the highest score for that set). For correct responses, we predicted that the number of correct responses would increase with the sophistication of the reporting tool, so we used the *SD* (34%) to test this. For the first analyses (overall correct responses), the *SD* was set to 6.08.

Results

Overall Correct Responses

The Bayes Factor between AICI and Free Recall and between Questionnaire CI and Free recall, indicated that the evidence substantially supported the alternative hypothesis $B_H = 182.55$ and $B_H = 16.65$ respectively; those between AICI and Basic Chat CI, between AICI and Questionnaire CI, and between Basic Chat CI and Free Recall were insensitive, $B_H = 2.54$, $B_H = 0.59$, and $B_H = 1.54$ respectively; and those between Questionnaire CI and Basic Chat CI substantially supported the null hypothesis, $B_H = 0.07$.

The Bayes Factors thus indicated that there was substantial evidence that Questionnaire CI and AICI elicited more correct items overall than Free Recall, even though only AICI did so *significantly*. They also indicated that there was substantial evidence to support the null (that there was no difference in the number of correct items) when it came to comparisons between Questionnaire CI and Basic Chat CI. Finally, more data were needed to explore the other comparisons. Therefore, while the significance testing indicated that there was no difference between AICI and Basic Chat CI, between AICI and Questionnaire CI, and between Basic Chat CI and Free Recall, the Bayes Factors indicated that the data did not support this conclusion.

Dialogue

For the dialogue items, we focused only on comparisons between the AICI and the other conditions, and the *SD* was set to 2.49 (34% of the highest score) for these comparisons. The Bayes Factors supported the null hypothesis when comparing AICI and Free Recall, $B_H = 0.24$, and AICI and Questionnaire CI, $B_H = 0.20$, but they were insensitive when comparing AICI and Basic Chat CI, $B_H = 0.69$ (inspection of fig 2 shows a mean score of 6.73 for Basic Chat CI users and 7.33 for AICI users). Thus, participants generally performed similarly in all conditions (compared to AICI), but more data were needed to compare the scores between AICI and Basic Chat CI. Thus, while

statistical analysis suggested that there was no difference between conditions, Bayes Factors suggest that when comparing the two chatbots, the data did not support this conclusion.

Action

Again, we focused only on comparisons between the AICI and the other conditions, and the *SD* was set to 1.15 (34% of the highest score) for these comparisons. The Bayes Factors supported the null hypothesis when comparing AICI and Free Recall, $B_H = 0.24$, and AICI and Questionnaire CI, $B_H = 0.24$. However, it supported the alternative hypothesis when comparing Basic Chat CI and AICI (inspection of fig 2 shows a mean score of 2.47 for Basic Chat CI users and 3.2 for AICI users), $B_H = 3.98$. The strength of evidence thus indicated that participants performed similarly for these comparisons, apart from when comparing the chatbots, as the evidence suggested that the Basic Chat CI elicited fewer action items than the AICI. Therefore, again the lack of significance when comparing chatbots cannot be interpreted as support for the null, as the Bayes Factor indicates that there was evidence that the AICI performed substantially better than the Basic Chat CI.

Facts

Again, we focused only on comparisons between the AICI and the other conditions, and the *SD* was set to 1.54 (34% of the highest score) for these comparisons. Inspection of fig. 3 revealed that the mean score for users of the AICI was lower than those for using the Questionnaire CI and the Basic Chat CI, so rather than the testing the hypothesis that AICI users would perform better than these conditions against the null (there would be no difference between conditions), we tested the strength of evidence of the size of the differences. The Bayes Factors indicated that there was substantial evidence that AICI also elicited fewer than Basic Chat CI and Questionnaire CI, $B_H = 814.10$ and $B_H = 115.47$ respectively. For the comparison between Free Recall and AICI, we

re-set H_1 to the original prediction. The Bayes Factor indicated that the results between Free Recall and AICI were insensitive, $B_H = 1.34$.

Basic Chat CI was therefore *significantly* better at eliciting factual items than AICI, but the Bayes Factor indicated that Questionnaire CI also elicited substantially more items than AICI. However, to evaluate the performance of AICI against Free Recall, more data was needed (inspection of fig 3 shows a mean score of 2.5 for Free Recall users and 3.1 for AICI users). Thus, it was not possible to conclude that there was no difference between these conditions.

Description

Again, we focused only on comparisons between the AICI and the other conditions, and the *SD* was set to 1.46 (34% of the highest score) for these comparisons.

The Bayes Factors supported the alternative hypothesis when comparing AICI and Free Recall, $B_H = 50214.61$; AICI and Questionnaire CI, $B_H = 3059.80$; and AICI and Basic Chat CI, $B_H = 57.14$. Therefore, in this case, the Bayes Factor supported the significant results for these comparisons.

Overall Incorrect Responses

For incorrect responses, we expected the number of mistakes to be *fewer* as the sophistication of the reporting tool improved (the *SD* was set to $x = 0.83$).

The Bayes Factor indicated that the results between Questionnaire CI and Basic Chat CI, $B_H = 11.99$, Questionnaire CI and AICI, $B_H = 296.03$, supported the alternative hypothesis. The chatbots elicited substantially fewer mistakes than the Questionnaire CI. However, when compared to Free Recall, participants using Questionnaire CI elicited substantially more mistakes, $B_H = 51.17$. Comparisons between AICI and Basic Chat CI, and Free Recall and Basic Chat CI were

insensitive, B_H = 1.99 and B_H = 1.54 respectively, while those between AICI and Free Recall supported the null hypothesis, B_H = 0.29.

Thus, while only participants in the Questionnaire CI condition elicited *significantly* more incorrect responses overall than Free Recall, the Bayes Factors indicated substantial evidence that Questionnaire CI also encouraged more incorrect responses than both chatbots. Bayes Factors allowed us to conclude that there was no difference in the number of incorrect responses between AICI and Free Recall, indicating that these two tools encouraged accuracy more than the other two. Finally, Bayes Factors indicated that we could not conclude that there were no differences between the Basic Chat CI and Free Recall, or between the Basic Chat CI and AICI.

Errors

For these analyses, we focused again on comparisons between the AICI and the other conditions, and the *SD* was set to 0.28 for these comparisons.

The Bayes Factor indicated that the results between Basic Chat CI and AICI, and Questionnaire CI and AICI supported the alternative hypothesis, B_H = 11.30 and B_H = 8.61 respectively, and those between Free Recall and AICI supported the null hypothesis, B_H = 0.26 (the *SD* was set to x = 0.28).

Therefore, while significance testing suggested that it made no difference which reporting tool participants used. Bayesian analysis indicated that participants using AICI made fewer errors than those using Questionnaire CI or Basic Chat CI, and that there was *no difference* in the number of errors made between AICI and Free recall.

Confabulations

For the final analyses, we also focused on comparisons between the AICI and the other conditions, and the *SD* was set to 0.57 for these comparisons.

The Bayes Factor indicated that the results between Free Recall and AICI, and between Questionnaire CI and AICI, $B_H = 4.30$, and $B_H = 106.08$ supported the alternative hypothesis respectively (performance improved as the sophistication of the tool increased). However, the comparison between Basic Chat CI and AICI, $B_H = 0.48$ was insensitive.

Therefore, while only Questionnaire CI encouraged participants to confabulate *significantly* more than Free Recall, Bayesian hypothesis testing indicated that it also encouraged participants to confabulate more than those using AICI. The results also suggested that, rather than being no difference between Basic Chat CI and AICI (inspection of fig 2 shows a mean score of 1.07 for Basic Chat CI users and 0.97 for AICI users), there was not enough data to make a conclusion either way.

Discussion

Statistical analyses indicated that the AICI elicited more correct responses without compromising on accuracy and that this chatbot was particularly good at eliciting descriptive details, but could improve on fact gathering. However, it failed to reveal nuances in the data that the Bayes Factors did.

We considered Bayesian hypothesis testing to be appropriate for this type of research, as the differences between conditions were chosen to be subtle, and the stimulus was a short video (1 minute 45 seconds) that was not able to elicit dramatic differences in the number of recalled items in actual terms, so we anticipated that Bayes Factors might clarify the results somewhat. We also wanted to test the minimum number of participants possible. Although we made power calculations to reach this number, as Bayes Factors do not rely on power calculations, we considered them to be suitable to clarify the results. They also confirmed in many instances that the number of participants that we had tested was sufficient.

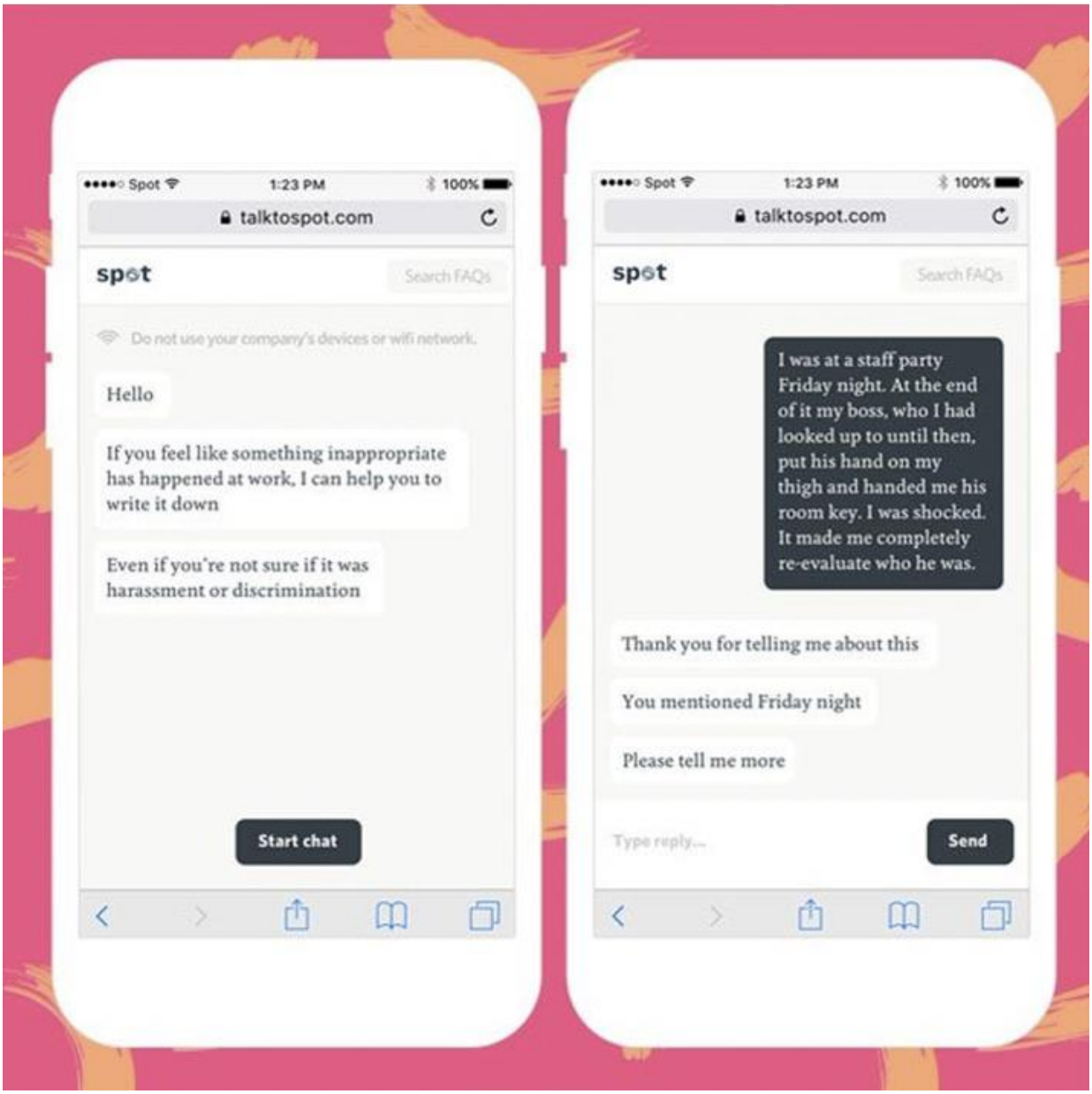
The Bayes Factors allowed us to make conclusions that were not possible when using statistics, and in some instances supported the statistics, adding weight to the implications. For instance, when it came to recalling correct information, significance testing indicated that overall, the AICI helped people to recall more items overall than Free Recall, that the AICI was better than the other conditions at eliciting description, while the Basic Chat CI was better than the AICI at fact gathering, and the Bayes Factors supported these results.

However, when it came to non-significance, statistical analysis fell short in three ways. An example or two will be given for each. First, the Bayes Factor indicated that while there was no significant difference in the number of action items between AICI and Basic Chat CI, the Bayesian evidence supported the notion that AICI elicited more items. Second, a non-significant result shows support for H_0, but in several cases, the Bayes Factors indicated that this conclusion could not be made. For instance, when comparing the number of dialogue items recalled by AICI and Basic Chat CI users significance testing indicated that there was no difference between conditions, but the Bayes Factors did not support this conclusion. Therefore, non-significance is *not* evidence in support of the null. Third, Bayes Factors were helpful when comparisons did show support for the null. For example, comparisons between Questionnaire CI and Basic Chat CI indicated that there was substantial support for the null in the number of correct items recalled overall, but orthodox statistics failed to demonstrate this.

Thus, by using Bayes Factors, we made more nuanced conclusions when it came to describing the effects, indicating that Bayesian hypothesis testing is a useful way of interpreting the results in studies of this kind.

Appendix 3

Example of AI Chatbot:



Mild harassment scenario link:

https://youtu.be/C0LMPW46EQs

Mild harassment scenario screenshot:

Example Transcript:

1. Please tell me everything you can remember about what happened. Try not to leave anything out, even if it seems trivial. I have as much time as you need.

2. Thank you for telling me about that. You mentioned… Please elaborate.

3. Thank you for telling me about that. Please provide specifics about the month, week, day, or time this happened.

4. You mentioned *an office*. Please describe.

5. You referred to *Mike*. Please tell me more about him or her.

6. The way this situation made you feel is important for understanding the impact on your wellbeing. Please describe in more detail how you felt **as you were experiencing the situation**.

7. How did it affect your wellbeing **after it happened**?

8. Did you tell anyone about about the event?

9. We're almost done. Before we finish, is there any other evidence of what happened? For example screenshots, emails, meeting notes, text messages, or recordings?