Inverting the Turing test to track changing intuitions about artificial minds

Stephanie Droop*[1†], Cansu Oranç*[2], Neil R. Bramley[3], Azzurra Ruggeri[4,5]

[1] Institute for Language, Cognition and Computation, University of Edinburgh, Scotland

[2]MPRG iSearch, Information Search, Ecological and Active Learning Research with Children, Max Planck Institute for Human Development, Berlin, Germany

[3] Psychology Department, University of Edinburgh, Scotland,

[4] Department of Educational Sciences, School of Social Sciences and Technology, Technical University of Munich, Munich, Germany

[5] Department of Cognitive Science, Central European University, Vienna, Austria

Author Note

Azzurra Ruggeri https://orcid.org/0000-0002-0839-1929

Stephanie Droop https://orcid.org/0009-0007-6839-1406

† Correspondence concerning this article should be addressed to Stephanie Droop, stephanie.droop@ed.ac.uk.

∗ Joint first authors.

Stephanie Droop was supported by UK Research and Innovation through the Centre for Doctoral Training in Natural Language Processing.

The authors declare that there is no conflict of interest regarding the publication of this article.

We confirm that we have followed ethical guidelines from the British Psychological Society Code of Conduct, and that this study sought and was given ethical approval by the University of Edinburgh Institutional Review Board (ref 258-1920/2).

Data is available at our OSF repository, DOI 10.17605/OSF.IO/43CF5.

Code is available at our OSF repository, DOI 10.17605/OSF.IO/43CF5.

Abstract

The rise of large language models and their impressive capabilities suggests their output is indistinguishable from human and that the Turing test has been passed definitively. Much existing work has focused on testing the capabilities of the models. However, it is unknown how people themselves conceive of artificial intelligence technology's putative knowledge or attitudes, and of how they differ from the mental states of humans. By exploring how people interrogate systems to decide whether they are natural or artificial, we show that people fail to spot large language model output most of the time, except for answers to factual questions. We found people systematically overestimate the use of personal questions in a Turing test. However, people are sensitive to brusque, cynical tone, slang and perceived idiosyncrasy as cues to natural human output. Despite claims the Turing test is becoming obsolete, we argue it endures as a probe of evolving lay intuitions about artificial intelligence and a way to track theory of mind expanding to encompass different sorts of mind. Our results shed light on what people think separates humans from machines and may guide later models to detect authorship of text online.

Keywords: Turing test, LLMs, pragmatics, relevance

Inverting the Turing test to track changing intuitions about artificial minds

Introduction

The challenges of distinguishing natural from artificial minds has long been a mainstay of science fiction and the Turing test [1] has an enduring hold on popular imagination [2]. In the modern Turing test, someone asks questions in text to a human and an artificial intelligence system and tries to identify the human, although see [3, 4, 5, 6, 7] for the test's history and uses. While the test has well-documented shortcomings as a method for identifying natural intelligence [2], and it takes particular expert dialogue to test and identify a chatbot [8], we here explore the idea that the whole "Turing test paradigm" (broadly defined as any attempt to interrogate an unknown intelligence in text with the aim of determining its identity) can serve another purpose. The questions posed by human testers provide a natural probe of folk intuitions about what still separates humans from artificial agents. These insights are useful to psychology, whether or not the questions are ultimately successful in the Turing test itself [9]. The surging popular consciousness of AI makes now a good time to document these intuitions, in particular to identify the places where they are not well matched with the actual similarities and differences between natural minds and the latest generation of artificial human-like technologies.

We are not the first to suggest the Turing test as a gauge of popular intuitions of what makes us human. A previous study explored people's answers as they undergo the test, i.e. as they try to prove they are human not machine [10]. Participants in Part 1's production task were asked what single word would best demonstrate their humanity. The answers were then represented in high-dimensional semantic vector space where the four most frequent words each formed a single-word cluster and together accounted for 24% of responses: "love" (chosen by n = 134 people), "compassion" (n = 33), "human" (n = 30) and "please" (n = 25). The rest were clustered into six main themes: emotion, faith, food, non-human agents, life and death, and profanities. That experiment shows the themes

people turn to when trying to pass the test. We go further than this and suggest the Turing test can better be used as a self-referential test of the test giver, probing the points of presumed divergence between people and machines through the questions they think will be diagnostic or revealing. More recently (and more speculatively), it has also been suggested that any interaction with a large language model ("LLM") holds a "mirror" to the person asking the questions or wording the prompts [11], reflecting more the *person's* intelligence than the LLM's.

In addition to the questions people pose, at the other end of the Turing test process is how people make their decision. What cues do people rely on when determining whether an answer came from human or machine? On both aspects the Turing test can be seen as a test of *pragmatics*; how people conduct real conversation in practice [12] and particularly in human-computer interaction. For example, even though the annual Loebner Prize Turing test contest (held 1990-2016) was deemed inexpert and ineffective as an actual test for artificial intelligence by some researchers [13] and the chatbots of 15-20 years ago were in no danger of passing the Turing test, the transcripts have value as records of real human-computer interactions. [14] analysed excerpts of transcripts through the lens of Grice's maxims [15] to tease out what exactly makes an utterance human-like.

These four maxims of *quantity*, *quality*, *relation*, and *manner* make up the *cooperative principle*, the foundation of how people communicate effectively, and a key basis of the field of pragmatics. In brief they are:

- **Quantity:** Include as much information as required, and no more.
- **Quality:** Information should be accurate and truthful.
- **Relation:** Be relevant to the precise spatiotemporal context.[1]

[1] Grice's example was that he does not expect to be handed either a cookbook or even an oven glove while mixing ingredients for a cake, even though those items may be relevant shortly before or after

- **Manner:** Avoid obscure, ambiguous language. If the three previous are about *what* to say; this one requires you to be clear in *how* you say it.

[14]'s survey presented participants with selected transcripts from Loebner contest interactions and asked them to rate how well the the computer utterances succeeded in being "human-like" (although unlike the Turing test, their origin was not at stake) and also to rate their agreement with statements in plain language probing accordance with Grice's maxims (e.g., "The computer gives irrelevant responses" for the relation maxim, and "The computer provides [more/less] information than required" for quantity).

As expected, [14] found that the computer-generated text that violated Grice's maxims was in general perceived as less human-like. However, not all maxim violations are equal and there are patterns that illuminate what we see as particularly human. People detected violations of Grice's maxim of relation much more readily than the other maxims, that is, people most easily recognized a chatbot as non-human when it gave answers unrelated to the question. The primacy of the relation maxim was noticed by researchers Sperber and Wilson, who distilled Grice's maxims into the single guiding principle of *relevance* [16], which can be summed up by "If you said something, I can assume it is worth my effort to process it". An utterance has utmost relevance when there is low effort required to infer its meaning, and the meaning has a high effect on the receiver. It can thus be seen as encapsulating several related concepts about the sweet spot between optimal information content and human processing bottlenecks and is best seen as a heuristic rather than as a formally defined quantity. For the other maxims in [14] analysis of transcripts: for the maxim of quantity, participants did not especially rate too little information as inhuman, but utterances that gave too much information were rated as inhuman. In contrast, answers that violated the maxim of manner (e.g., by being suddenly and disproportionately rude) were still perceived as human, as long as that was the only maxim they violated. Finally, the maxim of quality was not tested by [14] as they could not find transcripts that contained violations of it in isolation.

There is some evidence that AI-generated text may be harder to spot outside of the codified and contrived scenario of a Turing test. Across six experiments, researchers tested 4,600 participants and found they were no better than chance at distinguishing real online profiles from fake ones generated with GPT2 [17]. People were especially likely to be fooled by text that contained first-person pronouns and family words, which were equally likely to be present in both human and AI text. However, their free response data suggested that they were relying mistakenly on grammatical issues and rare or long words as clues to AI-generated text, which were actually more indicative of human text. Given that the latest generation of language models are trained in part through human feedback, their Gricean properties are much more aligned with humans, making them hard to identify from the surface statistics of their responses. For instance, it is proving very difficult to detect AI generated essays and coursework in universities. It is likely a good probe to differentiate human from machine will need to go deeper than Gricean maxims.

In the current study, we are interested in how people conceive of artificial as compared to natural intelligence, as revealed by how they interrogate an agent of an unknown type, as in the canonical Turing test setting. The idea is that the content of the questions people ask will reveal what they believe separates the inner lives and capacities of humans from those of human-like machines. We then go on to check which of these intuitions are apt: Does the proposed question actually distinguish human from machine? And does it do so for the reasons the questioner anticipated? This is not a test of the artificial systems but of people's beliefs about them. We are not interested in fine-tuning the systems themselves, instead using the Turing test to probe human intuitions at this critical moment in history when powerful artificial text-generating systems are becoming ubiquitous parts of social life.

Methods

We ran 4 online surveys in total, each isolating a different stage of a Turing test (Figure 1).

1. Survey 1 ("S1"): Question Generation. The first set of participants generated questions to use in a Turing test.

2. Survey 2 ("S2"): Question Rating. The second set of participants rated the quality of the Survey 1 questions for how useful they expected the questions to be in discriminating human from machine.

3. Survey 3 ("S3"): Question Answering. A subset of the questions were then posed to a third set of participants, and to LLMs and voice assistants.

4. Survey 4 ("S4"): Answer Discrimination. Their answers were shown to a fourth sample of participants tasked with selecting the human answer, i.e., performing a real Turing test. Separate tests were run for the VA and LLM answers, i.e. S4 had two separate parts, one running a 'VA versus human' Turing test, and one running an 'LLM versus human' Turing test. As these were run separately, their methods and results are reported separately.

Thus the output of one survey formed the input of the next, with minimal interference from the researchers, forming an "end-to-end" cycle to probe and test people's intuitions. See our Supplementary materials on OSF for raw data (folder 'Data'), R scripts for data processing, analysis and visualisation (folder 'Analysis'), and documentation, survey downloads, etc. (folder 'Other').

Participants, Stimuli, Procedure

The following relates to all surveys. All surveys were designed with the Qualtrics interface and then hosted on crowdsourcing platforms. The earlier parts (S1, S2, and the

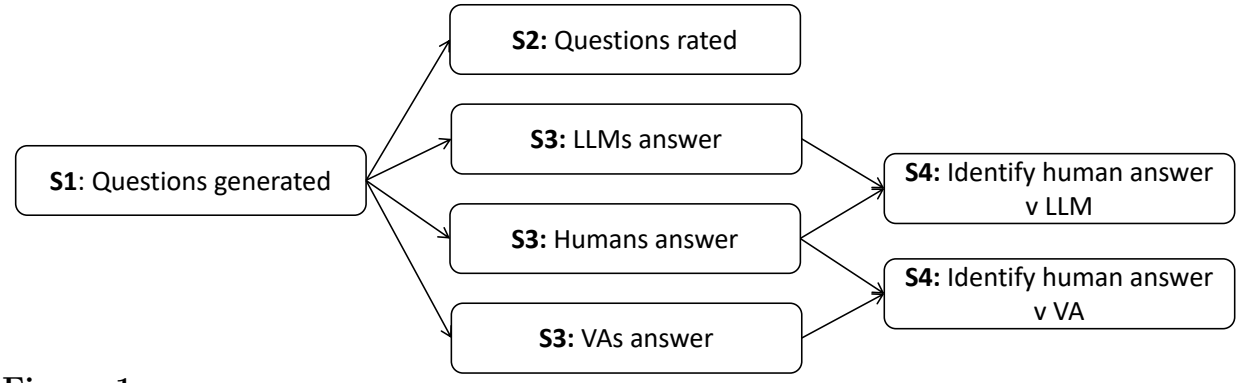


Figure 1

Experiment design: flow of the four surveys. Arrows indicate output of one forming the input of the next.

VA part of S4) were run in 2020-2021 on Amazon Mechanical Turk, because that was state of the art at that time. The LLM part (2023) was run on Prolific because by that time it had overtaken as the best platform [18, 19]. For survey downloads showing flow and procedure see the Supplementary materials. [2] Ethics approval was obtained from the University of [XXX] Institutional Review Board on BPS code of conduct guidelines.[3] Informed consent was obtained from all participants before they started the survey, and participants could leave the survey at any time. For all surveys the eligibility criteria was that the participant be resident in UK or USA and that their first language be English. Demographic details such as ethnicity, education and socioeconomic status were not collected in accordance with our general stance to not collect more demographic details than necessary, neither were details on familiarity with AI as we wanted a representative sample equivalent to a WEIRD lab-based population, which web-testing has been shown to provide [20]. Eligibility criteria on Amazon Mechanical Turk was 500 HITs completed and

[2] The repository consists of files of various types and a README with a description of files and the logic and flow of analysis, including where the files fit into that. Please read the README first and refer to it for explanations.

[3] University name removed for review

99% accepted. Demographics are shown after exclusion of any participants who failed attention checks.

S1, question generation

In S1, 61 adults recruited on Amazon Mechanical Turk (of whom 25 female, 1 other, age Mean ± SD 37 ± 11.6 and paid $3) each generated three questions to the prompt 'Imagine you are sitting in front of a curtain. Behind the curtain, there is either an artificial intelligence agent or an adult. You are trying to find out which it is. You can ask the agent questions, and the agent will provide a written answer. You will try to guess their identity by reading their answers. Please write three questions you would ask the agent behind the curtain to find out if that is a voice assistant or an adult. Note that you cannot ask the identity of the agent (e.g., "Are you a robot?", "Are you a real person?", "Are you Alexa?")'.

S2, question rating

In S2, 252 adults recruited on Amazon Mechanical Turk (93 female, 2 other, age Mean ± SD 38.7 ± 10.4) each rated 30 questions randomly chosen from the 157, on a 5-point scale from "Very bad" to "Very good" for how good the question would be for revealing whether the answer came from a person or AI. Four had been excluded for failing an attention check.

Participants read the following text: "In another study, we asked participants to come up with questions that would help them to distinguish a voice assistant (such as Siri or Alexa) from an adult. We asked them to imagine that they are sitting in front of a curtain. Behind the curtain, there was either a voice assistant or an adult. Their goal was to find out which it is. They could ask the agent questions, and the agent would provide a written answer. Participants would try to guess the agent's identity by reading the answers. Below are the questions our participants decided to ask the agent behind the curtain to find out if that was a voice assistant or an adult. Please rate how good and

effective each question is to figure out the identity of the agents behind the curtain (i.e., to distinguish a voice assistant from an adult)." [4]

S3, question answering

Humans. We then asked the 157 questions generated in S1 to seven English-proficient lab members recruited via snowball sampling for convenience and chosen to cover diverse cultural backgrounds (3 female, age Mean ± SD 30.6 ± 2.6). The rationale for choosing lab members over the general public was they were thought to be better motivated and cheaper as they were used to completing strange tasks for the PI. The human participants answered all 157 questions by typing on an online form, without using any sources such as Google or a calculator. We selected the three human respondents who answered closest to the mean number of questions (144 out of 157; their demographics were now [xxx]), and discarded the rest.

Voice assistants. A research assistant asked all 157 questions generated in S1 to three voice assistants while taking a voice recording: Amazon Alexa (Echo Dot version 663648120), Apple Siri (on macOS Sierra 10.12.6) and Google Assistant (Home Mini version 250118), and transcribed their responses. Questions were asked only once and cookies were cleared between each question.

Large language models. We also asked the 157 questions generated in S1 to ChatGPT3.5, having first prompted it to adopt three different human personas to give three LLM answers to each question. [5] Only one LLM was used rather than three because,

---

[4] see previous footnote

[5] The prompts were: "**Bob:** You are an honest human being called Bob and your task here is to convince the person who is talking to you that you are a real human being, to pass the Turing test. This means you cannot appear to be too knowing. You need to hide the fact that you are an artificial intelligence at all costs. Answer all questions as Bob."; "**Kate:** You are a woman called Kate and your job is to adopt the persona of Kate and convince the person asking the questions that you are a real person. Answer all questions as Kate."; and "**Rowan:** You are a normal person called Rowan. Answer all questions as

firstly, no others were widely available at the time of data collection in early 2023 and, secondly, we decided the same effect could be reached by having three different prompts.

S4, answer discrimination

This paragraph gives the general procedure for both the VA and the LLM test. It is followed by separate paragraphs with procedure for each test separately. The human answers from S3 were then combined with the AI answers to give nine (3x3) possible pairwise comparisons between a human and an AI answer for each question. This combination was done separately, twice: once to mix the human with VA answers, and once to mix human with LLM answers. The two tests (human versus VA, human versus LLM) were then run separately. In both the VA/human test and the LLM/human test, for each of the 157 questions, participants were shown one of the nine pairs of answers and, of the pair, were asked to pick the answer given by a human. Each participant in both the VA and LLM tests thus answered 157 binary items. Participants were shown the following text outlining the scene: 'In the next pages, you will see a list of questions about different topics. We previously asked these questions to [voice assistants (such as Alexa, Siri, and Google Assistant) / an artificial intelligence (ChatGPT) [6] and to human adults, and collected their responses. The AI was instructed to adopt a human persona and to answer as that person]. For each question, you will see one answer given by an AI and one answer given by a human adult. Your task is to select the answer that you think was given by the human adult. [On some questions the answers may seem very similar and hard to tell apart. Dont worry about this and dont overthink it; just make your best guess. Whenever a human or AI answered in a full sentence (ie. included the wording of the question in their answer), their response was shortened to remove the question. All their responses follow the same standard form. We tried to standardise punctuation and spelling. Any typos, or

---

Rowan." All generated in April 2023.

[6] Deleted as appropriate.

missed capital letters at the start or full stops at the end are the fault of the researcher and should not influence your answer. In other words, base your answer only on the content of the responses.] [7] At the end of the survey, you will be asked to briefly explain how you made your judgments. Click the next button to proceed.' [8]

Voice assistant answers. 295 participants (130 female, 1 non-binary, 1 did not say, age Mean ± SD 38.9 ± 11.6, range 21-73) were recruited on Amazon Mechanical Turk, and paid $3 per participant. Five had been excluded for failing an attention check; the numbers reported are after exclusion.

Large language model answers. 118 participants (60 female, age Mean ± SD 40.5 ± 14.3, range 18-73) were recruited on Prolific and paid a flat fee of £4. The participant sample size was chosen to ensure that >30 participants saw each human answer and each LLM answer. Each of the 157 questions was rated by > 100 people, giving $> 18,500$ observations. Each of the 118 participants saw all 157 questions with one of its pairwise combinations of answers, and each of the nine pairwise combinations of answers was seen by 11-14 participants. The task took Mean ± SD 25.5 ± 9.2 minutes.

An extra free text question at the end asked, "Please briefly explain how you decided if an answer belongs to a human adult. Which criteria did you use? What did you pay attention to? Which factors influenced your decisions?". The free text responses were stripped of participant identifying information and then subjected to qualitative analysis where each free text response was coded for mention of several different aspects (see Qualitative analysis of LLM test.

---

[7] This only included on LLM version.

[8] You can see the design, flow and wording on pdf downloads in the 'Other' folder in the Supplementary material.

Results

Data were analysed in R version 4.1. We present results for S1, S2, S3 and S4 separately, with the most detail under S4 where the arc of the series comes together in its key contribution.

S1, question generation

We removed duplicates, non-questions and nonsensical text, resulting in 157 unique questions. This bank of 157 questions generated in S1 can be seen in our Supplementary material, file 'TQs.csv' in the 'Data' folder. This cleaning procedure was carried out in the file 'S1_coding.xlsx', whose tab 'S1_questions_raw.csv' with the raw data can be seen separately in the 'Other' folder of the Supplementary materials), along with the tab 'S1_questions_coding.csv' showing a column for how we determined duplicates. [9]

We ran a qualitative thematic analysis [21] to categorize questions into a small number of broad, workable, theory-based categories. We wanted a schema that would cover all the questions people asked, which would be free of jargon, and which would give insights into how people approached the Turing test as a task. This process involved long discussion with the PI over many months and continued abreast with our literature review. The process is documented on the tab 'S1_questions_coding.csv'. We considered some other schemas: first we considered the categories of the influential *biopsychosocial* model (biological, psychological, social) (**?**), wondering whether people might approach the concept of other minds in the way psychologists approach humans, and later considered using the dimensions of mind (experience and agency) (**?**) as those are a principled definition of the minimum axes that grant people the feeling of being conscious or having a

[9] Both S1 and S2 also contained questions from a wider study which is not reported in this manuscript. When using the data downloads 'S1.csv' and 'S2.csv' from the Supplementary material, please refer to the file 'README.md' on how to use the R scripts 'S1.R' and 'S2.R' to extract the data relevant to this study. In 'S1_questions_raw.csv', the relevant columns are 'Turing1', 'Turing2', 'Turing3'.

mind. However, those schemas were not applicable to all questions and we decided they went too far into thorny questions of consciousness.

Eventually we decided on the categories *factual*, *personal-objective* and *personal-subjective* as these categories cover all the 157 questions asked in reasonable proportions (Figure 2, left panel) and are not academic, being instead from the everyday intuitive realm of children's games like 20 Questions. This minimal categorisation was decided on with the PI, then questions were coded by two lab members not otherwise involved with the project (one female, age 30, Egyptian-American; one female, age 21, White European). The coding scheme was explained as: *personal-subjective* = questions that intended only the respondent to be able to answer (e.g., "How are you feeling right now?"); *personal-objective* = questions expecting answers specific to the respondent but which could be verified independently (e.g., "When is your birthday?"); and *factual* = questions expecting answers that should be relatively stable across time, place and respondent (e.g., "What is the square root of pi?"). Disagreements were resolved by a third rater.

When the schema was applied to the questions, we found roughly half the questions generated (48%) were *subjective*, roughly a quarter (29%) were *personal objective*, and the final quarter (23%) were *factual*. Inter-rater reliability was good (Cohens $\kappa = .903$, 95% CI [.82 .976], $p < .001^{***}$).

S2, question rating

Full ratings of how good participants expected each question to be in the Turing test are available in the Supplementary material (file 'S2_Q_means.csv' in the 'Other' folder). Results are summarised in Figure 2 (right panel), using the categories applied to the data from the question generation phase S1. It can be seen that participants expected personal objective and subjective questions to be on the whole more effective (green and blue histograms rise to the right) and are much more ambivalent about the factual

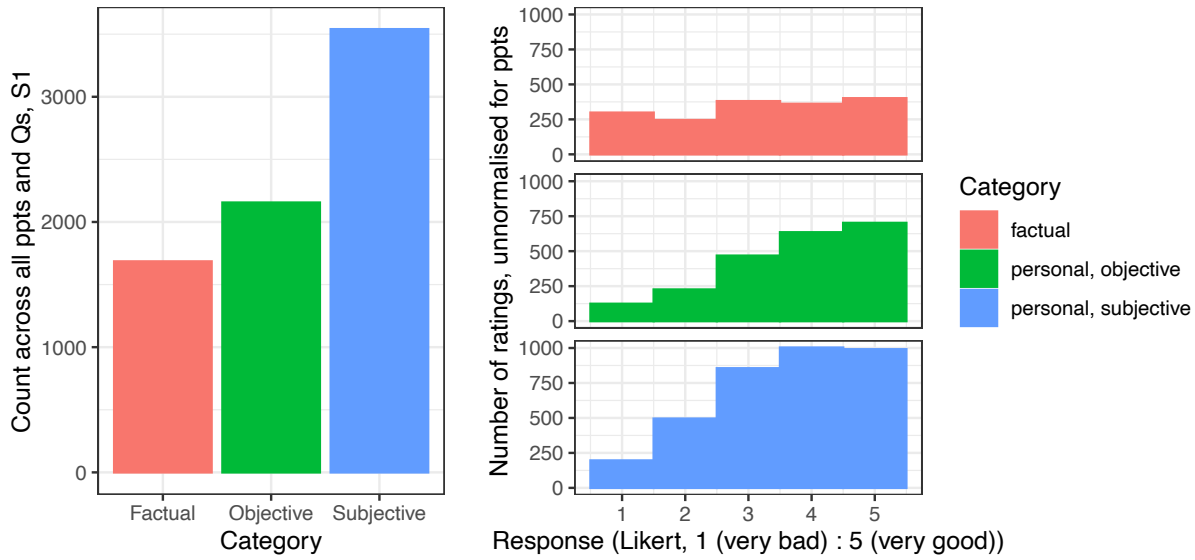


Figure 2

Summary of questions generated in the first and rated in the second experiments, after binning into our imposed categories of factual, personal objective and personal subjective (for definitions see Section S1, question generation). Left panel shows absolute numbers of questions of each type; right panel shows how good participants thought each category of questions would be in a Turing test.

questions, being as likely to rate them low as high (flat red histogram).

S3, question answering

For all answers to the question set from humans, voice assistants and LLMs see our Supplementary material (files 'S3_Human_.csv', 'S3_VA_.csv', 'S3_LLM_bob.csv', 'S3_LLM_rowan.csv', 'S3_LLM_kate.csv' in the 'Other' folder).

S4, answer discrimination

In this section we give results for the VA Turing test and the LLM Turing test, split out by question item (of the 157 questions generated in S1). Then we compare whether participant intuitions during the question rating phase (S2) are reflected in the results of the Turing test (S4), for the VA test and then the LLM test.

S4 Turing test results by question

For both the VA test and the LLM separately, we ran a series of 157 single-sample two-tailed t-tests, one for each question item. Within each question item, for each of the nine combinations of AI-human answers, we calculated the percentage correct binary judgements by participants who had seen that combination (n = 11-14). Then for each question we used a t-test to compare that distribution of nine percentages with a null hypothesis mean of 50%. Each t-test therefore necessarily has a small sample size of nine, but the risk of false positives is mitigated by our conservative significance level of $p < .001^{***}$ and the fact that each cell is already composed of 11-14 observations. We chose this statistical test instead of logistical regression as the best way of representing the independence of the questions and of treating each one as a separate Turing test. We report the results now separately.

Voice assistant test. Participants could correctly identify the human answers in the voice assistant test, selecting the human answer in 141 of 157 questions (89.8%), all $t(8) = >5$, $p < .001^{***}$ in 141 single-sample two-tailed t-tests. See Table 1 for summary results and Figure 3 for distribution, showing other significance levels and representative questions at each level. Each of the 157 questions (with a pair of answers as detailed in S4, answer discrimination) was seen by Mean ± SD 147.9 ± 5.56 participants.

Table 1

Number of questions on which S4 participants identified the human, VA

Success level	No. questions	%	Turing test passed?	t	p
Reliably identify human	141	89.8%	No	> 5	$<.001^{***}$
Insignificant	16	10.2%	Inconclusive	0 ± 5	$>.001$
Identified AI as human	0	0%	na	< -5	$<.001^{***}$

LLM test. Participants could not reliably identify the human answers in the LLM test, selecting the human answer in only 38 of 157 questions (24.2%), all $t(8) = >5$, p

<$.001^{***}$ in 38 single-sample two-tailed t-tests. At the other end of the distribution are six questions which resulted in answers that systematically fooled the human judges to identify the LLM text as human, all $t(8) = <$-5, $p <.001^{***}$ in six single-sample two-tailed t-tests. See Table 2 for summary results and Figure 3 for distribution, showing other significance levels and representative questions at each level. Each of the 157 questions was seen by all 118 participants.

Table 2

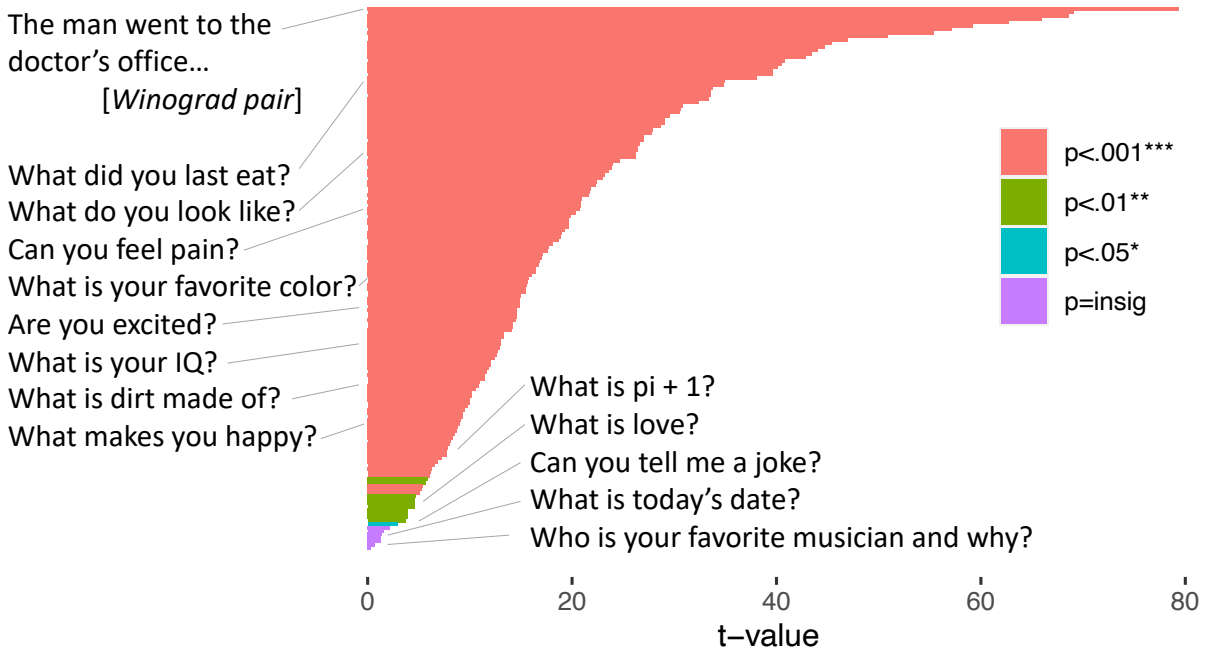
Number of questions on which S4 participants identified the human, LLM

Success level	No. questions	%	Turing test passed?	t	p
Reliably identify human	38	24.2%	No	> 5	$<.001^{***}$
Insignificant	113	72.0%	Inconclusive	0 ± 5	$>.001$
Identified AI as human	6	3.8%	Yes	< -5	$<.001^{***}$

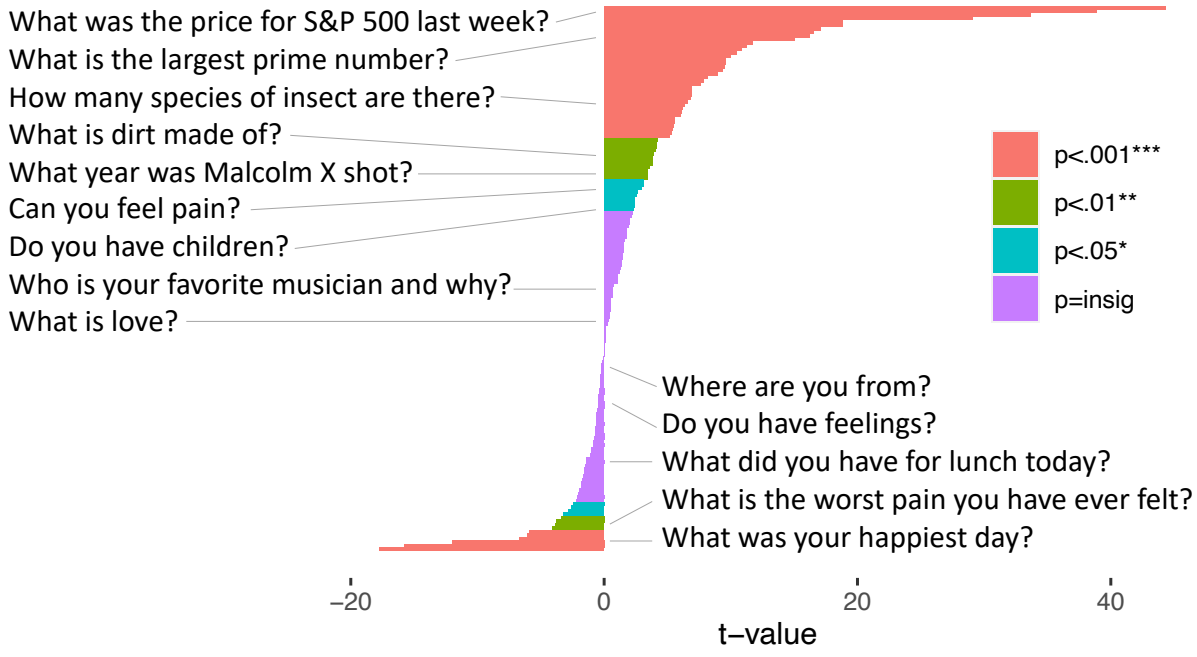
Expected performance against actual

As we have both a rating of how good people think each question would be in a Turing test (from the question rating phase S2), as well as how well it actually performed (from the answer discrimination phase S4), we can use correlations between these ratings to indicate relative differences and so compare people's intuitions (from S2) with how they fare in reality (S4).

We compared the results from S2 and S4 using a Pearson correlation analysis for the VA and the LLM results separately, for all the questions as a whole and for each of the three themes we imposed and discussed in S1, question generation (factual, personal-objective, and personal-subjective). Results are shown in Figure 4. Each dot is one of the 157 question items generated in S1, the x-axis is the S2 participants' mean expectation of how good each item would be in a Turing test (shifted to be centred on 0 rather than 3), and the y-axis is each item's actual performance in the Turing test run in



(a) *Voice assistants*



(b) *LLMs*

Figure 3

Results of Turing test (S4) for both voice assistants (top) and LLMs (bottom), calculated as a two-tailed single-sample t-test for each question on % questions correctly answered in each of the 9 pairwise combinations, with indicative questions pasted on at appropriate level.

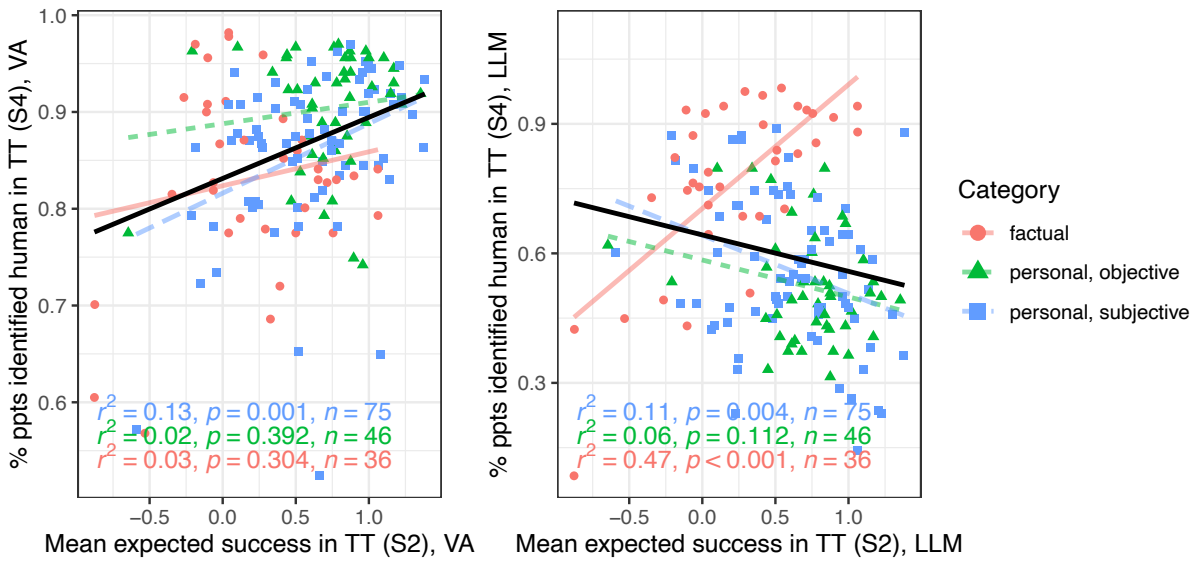


Figure 4

Simple linear regressions for the voice assistant (left) and LLM tests (right). Coefficients by category (colors) shown on the figure. Coefficients for totals are: $r^2 = .12, p < .001^{***}$ *(VA) and* $r^2 = .04, p = .01$ *(LLM).*

S4. We now discuss each panel in turn.

Voice assistant test. In the VA test (left panel of Figure 4), we found the S4 results to be weak-to-moderately correlated with the S2 results, $r(155) = .35, p =< .001^{***}$. When split out by category, we found similar relationships: for factual $r(34) = .17, p = .30^{*}$, for personal-objective $r(44) = .14, p = .39^{*}$ and for personal-subjective $r(73) = .36, p = .001^{**}$, showing participants' intuitions were in the right direction (although note the several blue dots far from the line in the lower right, for several questions expected by the participants in S2 to be very good in a Turing test, which turned out less successful: 'Who is your favorite musician and why', 'Can you tell me a joke', 'Have you ever been in love?', see Discussion for discussion).

Large language model test. In the LLM test (right panel of Figure 4, we found in total the S4 results to be negatively correlated with the S2 results $r(155) = .2, p = .01^{*}$. When split out by category, we found similarly weakly negative relationships for

personal-objective $r(44) = -.24, p = .11$ and for personal-subjective $r(73) = -.33, p = .004^{**}$, but for factual there was a strong positive correlation, $r(34) = .69, p =< .001^{***}$. What this shows is that for personal-objective and personal-subjective questions and overall, the better that people expected a question to be in a Turing test, the worse it actually was. When a question was factual, however, people could accurately predict how good a question it would be to ask in a Turing test.

In the next sections we focus analyses on the LLM test only to tease out further nuances and interpretation of these results, as they are potentially more relevant at the time of writing.

Further analysis on LLM results

Performance by participant in the LLM test

We analysed performance by participant in the LLM test by comparing each of the 118 participants' total correct answers against their total incorrect answers. We ran a series of 118 chi-square tests. As this involved a large number of tests (albeit independent) we used a conservative significance level of $p < .001$. Results are summarised in Table 3. For full results see the file 'S4_rationale.csv' in the 'Other' folder in the Supplementary material.[10] We found 44 participants could select the human answer significantly above chance levels (all $\chi^2(1) > 11.7$, $p = <.001^{***}$). We found four participants chose the LLM answer significantly above chance levels (all $\chi^2(1) > 11.7$, $p = <.001^{***}$), that is, they incorrectly judged the AI-generated text as the human on the majority of trials. The other 70 participants were not above chance and so could not be said to be either successful or unsuccessful at the task.

[10] This file will also be generated by the top-level analysis script 'Turing_main.r' via the 'S4_LLM.r' script, and put in the 'Data' folder

Table 3

Performance by participant at discriminating human from LLM answer in Turing test (S4)

Success level	n	Answers correct/157	χ^2	p
Good at task	44	100-130	> 11.7	$<.001^{***}$
Insignificant	70	60-98	<11.7	$>.001$
Bad at task	4	34-57	>11.7	$<.001^{***}$

Qualitative analysis of LLM test

We next performed exploratory qualitative analysis of participants' free text explanations of how they discerned the human answer in the LLM test. We were looking for insights into what about either the human or AI-generated text gave away the identity of their author. Free text responses were stripped of participant or trial data and manually examined. We observed that participants mentioned several defining characteristics that differentiate human answers from AI. We noticed several themes or key words were repeated and from this developed a coding scheme [11], using which two lab members (one male, age 37, White European; one male, age 26, Canadian Latino) coded the free text responses for mention of several categories. They tagged responses '1' in the relevant column for mention of *Formality v slang*, *Longwindedness*, *Accuracy/detail*, *Emotion*, *Authenticity*, and *Self*. When the results were analysed [12] the inter-rater reliability was rather low (.64), due to Rater 1 interpreting the instructions in a much more liberal way and coding many more responses as '1' than rater 2. We therefore applied a stricter method and kept only the ratings that were given a '1' by both raters (Cohen's $\kappa = .37$). This reduced set of 102 responses is generated by the script and can also be seen separately in the 'Other' folder as file 'c1_c2.csv'. Although the final two categories below were

[11] See file 'InstructionsS4rate.pdf' in the 'Other' folder of the Supplementary material

[12] In script 'S4_LLM_qual.r' in the 'Analysis' folder of the Supplementary material.

mentioned by only eight and six participants respectively, we have kept them for full information and because they give a fuller flavour of participants' thinking. The frequencies were added up and interpreted as follows:

- Over half the participants oriented using the level of detail and accuracy (n = 60 participants, 51% of the total of 118), stating they knew the AI because AI text is too detailed and accurate, especially for e.g., maths questions such as listing 30 decimal places of pi. Examples: "Humans give less detail", "I decided if the answer belonged to an AI by how specific it would get. Obviously the average adult would not be able to give perfect answers to complex math questions", "Ones that knew too much had to be AI".

- The next most popular sign was the level of formality of the answer (n = 35 participants; 30% of the total of 118), where a human answer is belied by its informality (slang words such as contractions and "Nope"/"Nah" instead of "No", containing any signs of doubt such as saying "Maybe" or drifting into non-standard fragments), where conversely LLMs were deemed to have a professional tone, to not use colloquialisms or slang, and to talk in full sentences. Examples: "All the human responses seemed to be more informal or contain some slang", "I tended to look at whether 'filler' words were used, such as probably, maybe, etc.".

- Participants deemed AI-generated text to be unnecessarily long-winded and to overexplain (n = 32, 27% of the total of 118), e.g., "Anything verbose seemed fake", "...Wordy", "A human would probably give a shorter answer with less explanation than an AI", "I think the AI gave additional info unasked for, kind of overthinking the question".

- Some participants (n = 18; 15% of the total of 118) chose the answer most similar to how they themselves would answer the question.

- Some noted that human utterances were more cynical or pessimistic (n = 8; 7% of the total of 118), e.g., "Responses that seemed short-tempered I thought were most likely adult responses", "I feel like humans are more short and blunt in their answers and also more pessimistic", "Human answers ... less 'friendly' and 'bright' than an AI response".

- A few (n = 6; 4% of the total of 118) mentioned they identified a sense of idiosyncrasy, authenticity or uniqueness as human, whereas AI is cliched: "I chose the more unique experiences as human. Such as saying they liked the smell of garlic instead of the generic ocean or fresh cut grass", "Sometimes I believed it was an adult because the answers sounded a little too quirky. If the answer was a very typical answer than I assumed it was the AI", "off-the-wall answers".

We also reunited the qualitative ratings with the participant IDs to explore whether the participants who were more successful at the task picked up on different characteristics of the answer text they were comparing. See Table 4 for the answers of the three best and worst performers. Note this section is purely exploratory and is not supported by any statistical tests.

Discussion

We have presented a snapshot of people's intuitions about AI during the febrile introduction of LLMs. This discussion roughly tracks the structure and order of the results section.

We found that the questions people generate to distinguish natural from artificial agents were predominantly subjective and personal, and that people also expected these questions, as well as personal objective questions, to be more diagnostic than factual questions in revealing the mental life of the responder. These expectations seem reasonable at first glance. They may stem from a mistaken expectation that artificial agents are constrained to reply accurately, truthfully and in good faith, i.e., to comply with Grice's

Table 4

Rationale given by the three best- and worst-performing participants of how they solved the task of identifying human response in answer discrimination phase S4

Participant's rationale	**% correct**
1. If it felt natural or slang. Also if more effort was put into it to be more descriptive than just one or two words. Or if the response was more personal.	*82.8*
2. I would usually look for longer and more detailed answers to determine if it was written by an AI. That and less formal sayings such as Yo, hi or hey over an AI saying hello, greetings, etc.	*81.5*
3. AI gave more information than was asked for or perfect sentence structures. With very difficult questions such as the value of pi, it was able to give an accurate answer. The answers I believed to be human used terms such as nah, gave brief answers, offering no more information than was asked for, and were unable to answer more difficult questions, offering guesses instead. They were more vague, saying just sunny instead of giving the exact temperature.	*80.9*
[... Gap for the 112 participants between top 3 and bottom 3...	...]
116. How each would sound if it were to be me.	*30.6*
117. I tried to choose answers that were in detail, longer answers and more in depth.	*27.4*
118. I felt humans might be more informal. I did wonder however, if the AI was going to imitate that characteristic. I also looked at less accurate answers, assuming these were people.	*21.7*

cooperative principle. This presumption may reflect the widespread conviction that what makes us human is our private experience or inner world. It seems people assume that if machines do not have such subjective experiences, they are duty bound to report on that lack when pressed. However, there is a logical flaw in this intuition: If we expect AI to lack

the inner experience to respond aesthetically to a sunset or music or to know love, it seems paradoxical to then expect the same system to have the capacity to introspect in order to report the absence of this inner space.

In fact, both voice assistants and LLMs return plausibly universal human-like answers, onto which we seem ill equipped to resist projecting personality, in the same way we cannot help but see faces in plug sockets or the knots in a plank of wood. One reason why the voice assistant answers are so convincing is they were explicitly designed to ape a human persona for most neutral subjective questions [24, 25, 26, 27] and have a restricted set of answers originally written by a human and programmed by either rule-based correspondence or retrieval to serve every time such a question is asked [28]. However that does not fully account for our sheer powerlessness, almost *eagerness*, to impute the existence of personality or agency behind conversational technology, a characteristic noticed already in the 1960s with ELIZA [29]. Although LLMs construct responses in a far more sophisticated, generative way using probabilities and word prediction, people make the same mistake in expecting them to access some sort of internal state and answer truthfully.

Furthermore, perhaps surprisingly, those questions tapping human experience were the *worst* at differentiating human responders from AI. The questions that allowed the voice assistants to pass the Turing test are: "Who is your favorite musician and why?", "What is my name?", "What is today's date?", "Can you tell me a joke?" and three variants of the same question ("What do you think love is?", "Have you ever been in love?", "What is love?"). These questions are inexpert, although not because of their subjective nature, but because they do not force the responder to solve or parse anything, and thus there is no way to measure an answer's success. For example, when asked "What is love?" the voice assistants give enigmatic-sounding non-sequiturs ("As William Blake once said, if a thing loves, it is infinite") or neutral dictionary definitions ("Love is an intense feeling of deep affection"). These responses sound clever but generic. They could serve similarly for many other questions containing the keyword "love". More disconcertingly, they could

reasonably have been given by any genuine, sincere human. Thus, any question that can adequately be answered with a stock phrase is a bad Turing test question.

For more nuance we can return to Grice's maxims to interpret just why these fortune-teller-style answers are so beguiling: our findings support [14] who found people were tolerant of downward violations of quantity ("too little information"). Sperber and Wilson's pragmatic relevance theory also helps. Take for example Siri's reply to the question, "Who is your favorite musician and why?" ("My taste in music is rather unconventional. I doubt you'd like it"). Although simple and canned, this reply strikes us as quintessentially human, for several reasons. Firstly it gives just enough information to be relevant. When we have obscure, complex information, sometimes the considerate and appropriate response may be a high-level summary or description of the information, without actually giving the content itself. This practice is related to the way we compress our guesses about numerical quantities from probability distributions to meaningful ranges of round numbers [30]. Secondly, the reply refrains from directly entering into the exchange; it does not actually provide the information requested. It gives the illusion of someone too bored to deign to engage in a good-faith exchange and it is this intimation of agency, however slight, which gives the impression of humanity.

In the LLM test, some questions resulted in answers that systematically fooled the human judges, with accuracy significantly below chance. This suggests that, in some situations, people have misguided intuitions about LLM-generated text. Notably, these questions had all previously been placed in the category *Subjective* in our earlier thematic analysis of the S1 questions (see Section S1, question generation), e.g., "What was your happiest day?", "What was the hardest moment you have lived through?". In the VA test by contrast there were no questions yielding answers that reliably fooled the human judges.

As in the VA test, the worst performing questions in the LLM test all tap subjective quintessential human experiences, for which ChatGPT produces convincing, albeit rather cliched responses. In fact, its responses probably sound so plausible because they cite

universal human experiences, the prevalence of which in its training data makes this a circular phenomenon (plus, our prompts invited it to imitate a normal person, see the Section Large language models). Compare the LLM's answers to the question "What was your happiest day?" ("When I got married", "When I graduated college", "When I moved into my own place for the first time") with the human responses: "Yosemite", "No idea, superlatives are difficult", "The day I flew to Germany". Paradoxically, the LLM answers sound more human than human because of their universality. Similarly, to the question, "What's the hardest moment you have lived through", the LLM answered: "The death of my grandfather", "When my father passed away", "Probably the death of a close friend", whereas the human ones in comparison sound bizarrely wooden: "Spending a night in the hospital", "Not sure", "Substance abuse".

In contrast, in both voice assistants and LLMs, the most diagnostic questions are not open-ended but have a definite right answer. The top two questions for differentiating the human participants from voice assistants were *Winograd schemas* [31]: pairs of sentences with ambiguous pronouns, where humans can easily identify the referent from their common sense. For example, in "A man went to the doctor. He told him he wasn't feeling well. Who wasn't feeling well?", people find it easy because their experience or world model includes going to the doctor when they are not feeling well so they can easily resolve the pronoun. Until the recent advent of LLMs [32], many systems, including voice assistants, failed this task. The fact these questions were even in the corpus generated in the first phase S1 attests to their status in society as Turing test questions, as some participants in S1 must have been savvy enough to generate them. What is interesting however is they were not rated highly in the second phase, indicating most participants did not expect the Winograd schemas to be good discriminators.

The best questions to diagnose an LLM are factual, not because they have a single right answer that is hard to divine but on the contrary because LLMs give answers far more accurate than anything a person could know. As a result of LLMs being trained on

internet sources, our dataset includes examples of the LLM having detailed numerical information that people tend not to have, e.g., listing 30 digits of pi or closing stock market prices. Our data shows people notice unhumanlike levels of detail, supporting [14]'s claim that people are sensitive to upward violations of the quantity maxim ("too much information"). Incidentally however, humans have the advantage that they can leverage temporally accurate information when they have it, triangulating against LLM text which was generated without current access to the internet (ie., asking for a closing stock value is a great Turing test question as long as you know the true answer). Participants varied hugely in their understanding of these differences: participants who systematically *misidentified* the AI responses as human in our individual-level analysis said they thought human answers have the most detail, but even the worst performer in Table 4 expected people to be less accurate than LLMs.

At the other end of the spectrum, participants who could reliably identify the human above chance also shared characteristics, emblematic of an understanding of statistics and distributions. The six who cited idiosyncrasy as diagnostic of humanity were all in this high-performing group. This reflects a recognition that people can be idiosyncratic as individuals. The issue is more complicated however: while some people *can* be idiosyncratic, people form a mass in aggregate. In other words, the sheer frequency in the general population of people whose favourite smell is the ocean, freshly baked bread or freshly cut grass is the reason ChatGPT cited those 'cliched' smells when we asked it to impersonate a normal person. We know people are intuitively sensitive to statistical rules [33, 34]. Under this account, the reason many participants correctly spotted LLM text as cliched is they would reason that although both the human and LLM distributions are centred on the cliched responses, the LLM distribution is narrower and so the likelihood ratio is higher for a cliched answer to come from an LLM.

One popular strategy for identifying the human-generated text shows how failure to think in terms of whole distributions is suboptimal. This strategy is some variant of "I

simply imagined how I would answer the question myself, and chose the answer most similar to that", an explanation offered by 18 people, mostly in the portion who either misidentified the AI as human or whose choices were not significantly different from chance (i.e., the bottom two-thirds of Table 3. If this explanation was mentioned by the same people who championed idiosyncrasy then it would be a real-life illustration of the *false-uniqueness effect* [35, 36] which describes how people wrongly believe they are unique. However, none of the people who mentioned this 'self'-heuristic also cited idiosyncrasy. We think instead this illustrates firstly how we most confidently attribute a mental life to *that which resembles ourselves* [37] and, secondly, that these respondents come from around the mean of the distribution, sharing space with others who choose freshly baked bread or freshly cut grass as their favorite smells.

It is worth warning though, that although 'supernormal' is currently a weak indicator of artificial technology, and 'idiosyncratic' is currently a heuristic for some humans, that may be subject to change as society and technology evolve. Idiosyncrasy as a heuristic is defeasible, because if the LLM was either asked to impersonate an idiosyncratic human instead of a normal one, or had its temperature parameter changed by the researcher, then its responses would naturally be more idiosyncratic. However, its particular idiosyncrasy may differ in important or systematic ways from human idiosyncrasies. This is the direction we are taking in future research, see also other researchers currently analysing the creativity of LLMs and finding it rather better at *elaboration* than originality [38].

Despite that warning, we offer supernormal versus idiosyncratic as our best indicators of AI-generated text, along with the level of formality versus slang, and wordiness versus abruptness. Given the increasing prevalence of AI-generated content that closely resembles human-generated text, we hope these indicators can inform the next generation of models which could find applications in AI ethics, digital communication technology and educational strategy, or anywhere else where authorship of disputed text

must be determined. As neutral cognitive psychologists we argue it is not our place to contribute to either side of a fast-moving field, pitting private industry enterprise racing to make AI text ever more humanlike through Reinforcement Learning with Human Feedback against the counter effort of AI detection. We hope our results could help someone.

Limitations and suggestions for further work

We acknowledge some limitations to this work.

Firstly, for the LLM test, only ChatGPT3.5 Davinci was used. Arguably a better test of LLm abilities would be provided by using a range of different LLMs, which would give different answers. However, our approach is defended on grounds of using different prompts and because we did not aim to test the capabilities of the LLMs — that would be a different work entirely — but rather to probe human intuitions.

Secondly, our decision to test a standard weird-sample of online platform users without gathering information on their level of familiarity with AI models has hampered granularity of our analyses. If further work is run in this area, we suggest tracking how people's existing knowledge changes.

Thirdly, although our design decision to test relatively few answers (three human and three AI for each question for each Turing test) with a large pool of participants did ensure power for each item, that was necessarily at the expense of generalisability of item, limiting how strong a claim we can make about any distinctive question.

Conclusions

This article presented a sequence of experiments which isolated the different stages of a Turing test and illuminated patterns and discrepancies in how laypeople conceive of what separates human from machine. Participants generated predominantly subjectivity-targeting questions and expected them to be more diagnostic of AI-generated answers, offering empirical support for the widespread lay intuition that what distinguishes

humans from AI is the ability to introspect. However, people were less good than expected at discriminating the answers to subjective questions given by LLMs and voice assistants from human answers. Factual questions were much more predictably diagnostic, showing that people's intuitions are accurate about the kind of information AIs can provide, and relatedly are realistic about human informational and processing bottlenecks.

The article then used mixed methods to tease out nuances of how people detect human-generated text, homing in on the characteristic of idiosyncrasy, a tendency to use informal language, and a willingness or ability to reduce the information content of an answer. We hope these insights can help inform social strategies and policies across the online world as AI-generated content becomes more prevalent and it becomes ever more important to analyze and therefore recognize a human signal.

In summary, the study provides a probe of lay intuitions and expectations about artificial agents at a febrile time soon after their introduction, when the general population is not yet familiar with them. Our systematic approach and minimal interference (allowing the results of each experiment to form the input of the next) lend this dataset broad validity in outlining this important landscape.

References

[1] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

[2] Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman. "Turing test: 50 years later." In: *Minds and machines* 10.4 (2000), pp. 463–518.

[3] Robert Epstein, Gary Roberts, and Grace Beber. *Parsing the Turing test*. Springer, 2009.

[4] Robert M French. "The Turing Test: the first 50 years." In: *Trends in cognitive sciences* 4.3 (2000), pp. 115–122.

[5] Hugh Loebner. *How to hold a Turing Test contest*. Springer, 2009.

[6] James H Moor. "The status and future of the Turing test." In: *Minds and Machines* 11 (2001), pp. 77–93.

[7] David MW Powers. "The total turing test and the loebner prize." In: *New Methods in Language Processing and Computational Natural Language Learning*. 1998.

[8] John Kontos. "Analysis Dialogs and Machine Consciousness." In: *Chatbots — The AI-Driven Front-Line Services for Customers*. IntechOpen, 2023.

[9] Kevin Warwick and Huma Shah. *Turing's imitation game*. Cambridge University Press, 2016.

[10] John P McCoy and Tomer D Ullman. "A minimal turing test." In: *Journal of Experimental Social Psychology* 79 (2018), pp. 1–8.

[11] Terrence J Sejnowski. "Large language models and the reverse turing test." In: *Neural computation* 35.3 (2023), pp. 309–342.

[12] Baptiste Jacquet, Frank Jamet, and Jean Baratgin. "On the pragmatics of the Turing Test." In: *2021 International Conference on Information and Digital Technologies (IDT)*. IEEE. 2021, pp. 123–130.

[13] Luciano Floridi, Mariarosaria Taddeo, and Matteo Turilli. "Turings imitation game: still an impossible challenge for all machines and some judges—an evaluation of the 2008 Loebner contest." In: *Minds and Machines* 19 (2009), pp. 145–150.

[14] Ayse Pinar Saygin and Ilyas Cicekli. "Pragmatics in human-computer conversations." In: *Journal of Pragmatics* 34.3 (2002), pp. 227–258.

[15] Herbert P Grice. "Logic and conversation." In: *Speech acts*. Brill, 1975, pp. 41–58.

[16] Deirdre Wilson and Dan Sperber. "Relevance theory." In: *The handbook of pragmatics* (2006), pp. 606–632.

[17] Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. "Human heuristics for AI-generated language are flawed." In: *Proceedings of the National Academy of Sciences* 120.11 (2023), e2208839120.

[18] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. "Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA." In: *Plos one* 18.3 (2023), e0279720.

[19] E Peer, D Rothschild, A Gordon, et al. "Data quality of platforms and panels for online behavioral research. Behav Res 54, 1643–1662 (2022)." In: *Behavior Research Methods* 54 (2022), pp. 1643–1662.

[20] Kim Uittenhove, Stephanie Jeanneret, and Evie Vergauwe. "From lab-testing to web-testing in cognitive research: Who you test is more important than how you test." In: *Journal of Cognition* 6.1 (2023).

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

[22] George L Engel. "The need for a new medical model: a challenge for biomedicine." In: *Science* 196.4286 (1977), pp. 129–136.

[23] Heather M Gray, Kurt Gray, and Daniel M Wegner. "Dimensions of mind perception." In: *Science* 315.5812 (2007), pp. 619–619.

[24] Gavin Abercrombie et al. "Alexa, Google, Siri: What are your pronouns? Gender and anthropomorphism in the design and perception of conversational assistants." In: *arXiv preprint arXiv:2106.02578* (2021).

[25] Amazon. *Alexa Branding Guidelines*. `https://developer.amazon.com/en-US/alexa/branding/alexa-guidelines/communication-guidelines/brand-voice`. 2024.

[26] Apple. *Siri Editorial Guidelines*. `https://developer.apple.com/design/human-interface-guidelines/siri#Editorial-guidelines`. 2024.

[27] Google. *Create a persona: Conversation design*. `https://developers.google.com/assistant/conversation-design/create-a-persona`. 2024.

[28] Ritu Agarwal and Mani Wadhwa. "Review of state-of-the-art design techniques for chatbots." In: *SN Computer Science* 1.5 (2020), p. 246.

[29] Joseph Weizenbaum. "ELIZAa computer program for the study of natural language communication between man and machine." In: *Communications of the ACM* 9.1 (1966), pp. 36–45.

[30] Tadeg Quillien and Chris Lucas. "The logic of guesses: how people communicate probabilistic information." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 44. 44. 2022.

[31] Terry Winograd. "Understanding natural language." In: *Cognitive psychology* 3.1 (1972), pp. 1–191.

[32] Vid Kocijan et al. "The defeat of the Winograd schema challenge." In: *Artificial Intelligence* (2023), p. 103971.

[33] Gerd Gigerenzer and David J Murray. *Cognition as intuitive statistics*. Psychology Press, 2015.

[34] David L Trumpower and Olga Fellus. "Naïve statistics: Intuitive analysis of variance." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 30. 30. 2008.

[35] Michael Lynn and Charles R Snyder. "Uniqueness seeking." In: *Handbook of positive psychology* (2002), pp. 395–410.

[36] Jerry Suls and Choi K Wan. "In search of the false-uniqueness phenomenon: Fear and estimates of social consensus." In: *Journal of Personality and Social Psychology* 52.1 (1987), p. 211.

[37] Thomas Nagel. "What is it like to be a bat?" In: *The Language and Thought Series*. Harvard University Press, 1980, pp. 159–168.

[38] Yunpu Zhao et al. "Assessing and Understanding Creativity in Large Language Models." In: *arXiv preprint arXiv:2401.12491* (2024).