On the Pragmatics of the Turing Test

Baptiste Jacquet
Laboratoire Cognition Humaine et Artificielle (CHArt-UP8)
Paris, France
& Université Paris 8
Saint-Denis, France
& Association P-A-R-I-S
Paris, France
baptiste.jacquet@paris-reasoning.eu

Frank Jamet
Laboratoire Cognition Humaine et Artificielle (CHArt-UP8)
Paris, France
& CY Cergy-Paris Université
ESPE de Versailles
Paris, France
& Association P-A-R-I-S
Paris, France
frank.jamet@paris-reasoning.eu

Jean Baratgin
Laboratoire Cognition Humaine et Artificielle (CHArt-UP8)
Paris, France
& Université Paris 8
Saint-Denis, France
& Association P-A-R-I-S
Paris, France
jean.baratgin@paris-reasoning.eu
Corresponding author

2021 International Conference on Information and Digital Technologies (IDT) | 978-1-6654-3692-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/IDT52577.2021.9497570

***Abstract*—The Turing Test was initially suggested as a way to give an answer to the question "Can machines think". Since then, it has been heavily criticized by philosophers and computer scientists both as irrelevant, or simply inefficient in order to evaluate a machine's intelligence. But while arguments against it certainly highlight some of the test's flaws, they also reveal the confusion that exists between thinking and intelligence. While we will not attempt here to define the concept of intelligence, we will instead show that such a definition becomes irrelevant if the Turing Test is instead considered to be a test of the humanness of a conversational partner instead, an experimental paradigm that can be used in order to investigate human inferences and expectations. We will review studies which use the Turing Test this way, not only in computer sciences where it is commonly used to evaluate the humanness of a chatbot but also its uses in the field of psychology where it can be used to understand human reasoning in conversation either with a chatbot or with another human.**

***Index Terms*—Turing Test, Chatbots, Cognitive Psychology, Pragmatics, Theory of Mind**

I. Introduction

With the recent advances of Artificial Intelligence, the idea that machines might become able to think for themselves sooner or later is making its way in the general population, helped by many movies like Spike Jonze's HER, Alex Garland's Ex Machina or Ridley Scott's Blade Runner among many others. But how close to this are we really? Can a machine really think? Do we have the tools necessary to evaluate it like Alan Turing [1] claimed with his famous Imitation Game or Turing Test (written TT from now on)?

This paper attempts to review some of the literature exploring the main issue that chatbots (the programs that can be evaluated by the TT) still face today: relevance, in other words, the ability to produce sentences that take into account the expectations of the users. Indeed, most of these programs still fail to sustain conversations of more than a couple successive sentences without relying on generic replies.

We will first describe what types of chatbots exist today and some of the technologies supporting them, before moving on to the ways that are used to evaluate them. We will then describe the TT along with some of its criticisms and the importance of conversational pragmatics in order to successfuly pass it. Finally we will review review some of the literature exploring the use of the TT as an experimental paradigm to study human reasoning and conversational pragmatics.

II. Chatbots

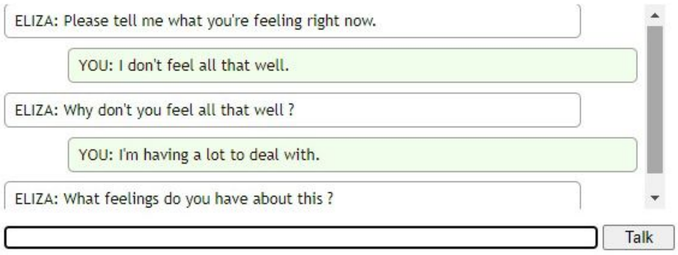
Virtual conversational agents, often called chatbots (or chatterbots), stem from an idea initially suggested by Turing in his test for intelligence [1]. They are machines which process text messages, attempt to find a relevant answer before sending a text reply to the user.

Chatbots are all but new, with the first application meeting this definition being ELIZA [2]. Meant to behave like a psychotherapist focusing on empathy and repetitions, this program uses keywords in the user's messages to send questions back to them. The conversation is kept going by the chatbot giving the impression of trying to dig dipper into the user's mind and psyche. It does so by analyzing the content of its input text, giving each recognized word a value before reorganizing the order of the words according to predefined patterns to generate a reply in the form of a question. If the content of the input doesn't match a recognized pattern, predefined fallback answers are sent back. ELIZA was initially designed to show how superficial Human-Machine conversation was, and yet users still ended up sharing deep and private aspects of their lives with it, often being fooled and refusing to believe they had been talking to a machine. A simple conversation with a version of ELIZA is shown in Fig. 1.

ALICE [3] is another chatbot which generalizes the concept of pattern matching to send replies to the user without necessarily asking questions back like a psychotherapist might. Just like ELIZA, it still uses keywords to detect patterns in the input. Below is an example of a simple exchange with ALICE asking about her favorite color.

Human: Hello ALICE.
A.L.I.C.E: Hi there!
Human: What is your favorite color?

This paper was funded by the Carnot Institute, for Project LAC (Langages Alternatifs et Cognition).



ELIZA: Please tell me what you're feeling right now.

YOU: I don't feel all that well.

ELIZA: Why don't you feel all that well ?

YOU: I'm having a lot to deal with.

ELIZA: What feelings do you have about this ?

Fig. 1. Example of a conversation we had with ELIZA. This version of the chatbot can be found at https://www.eclecticenergies.com/psyche/eliza

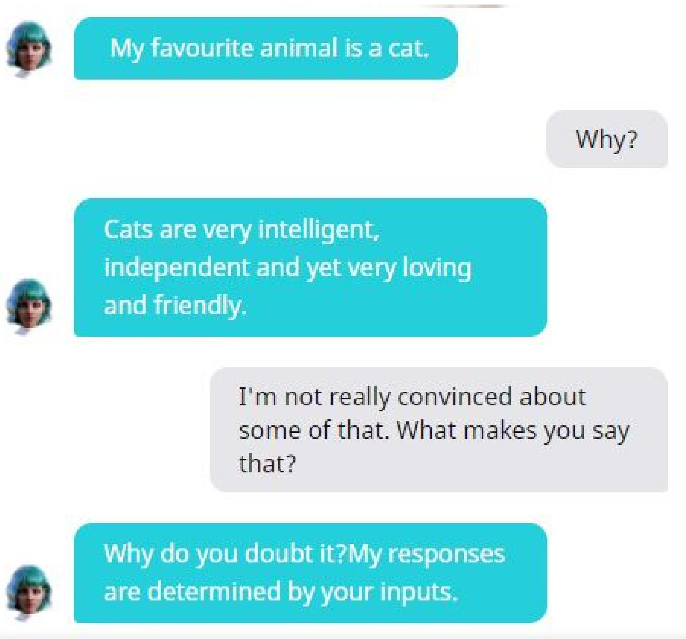
A.L.I.C.E: I don't have a favorite colour. But my favorite book is Don't Read Me.
Human: What can you tell me about it?
A.L.I.C.E: Ambiguous: "Tell me about it." – colloquial or "Tell me about ".

We can see above that ALICE also uses humor in this conversation, making it seem a bit more human, although it then fails to recognize the input in the last message sent by the user in our example. ALICE is built around its own language called AIML (Artificial Intelligence Markup Language) which is today used by many chatbots for its simplicity, making ALICE a good example or what can be expected of chatbots today.

While the simplicity and predictability of pattern-matching chatbots makes them quite attractive, even the most advanced of these types of chatbots cannot (to this day) deal with implicit content in the input messages [4]. Indeed, as we can see in the previous conversation with ALICE, it fails to understand what the user is referring to in the sentence "What can you tell me about it?". Mitsuku (now called Kuki.AI), the latest chatbot in this category to have won the Loebner prize (international competition based on the TT to evaluate chatbots), is still incapable of responding in a relevant way to implicit content in messages unless follow-up questions are pre-programmed (like about favorite color, pets..., see in Fig. 2) [4].

Pattern-matching is not the only tool available to create chatbots though. With the rise of machine-learning techniques, including supervised and unsupervised learning algorithms (depending on whether they use human-labeled data or not), new opportunities for the development of chatbots are being created, potentially allowing chatbots to use contextual cues more efficiently. Unfortunately no chatbot based on these techniques is currently openly accessible in order to test their actual abilities, though platforms exist to create them such as IBM Watson, Amazon Lex, Facebook Messenger, Microsoft LUIS and others. They often use algorithms such as LSTM (long short term memory) [5], RNN (recurrent neural networks) [6], BRNN (bidirectional recurrent neural networks) [7] or CNTN (convolutional neural tensor networks) [8].

There are two main types of chatbots based on these machine-learning techniques: retrieval-based models and generative models. Retrieval-based models read the input sentence



My favourite animal is a cat.

Why?

Cats are very intelligent, independent and yet very loving and friendly.

I'm not really convinced about some of that. What makes you say that?

Why do you doubt it?My responses are determined by your inputs.

Fig. 2. Example of a conversation we had with Kuki.ai. Notice how it answers the first follow up question correctly, but fails with the second follow up question. This version of the chatbot can be found at https://chat.kuki.ai/

produced by the user to create a thought vector representing the meaning of the sentence (or intent). This thought vector is then compared to entries in a database containing the possible answers the chatbot can give. The entry that is closest to the generated thought vector is selected and sent to the user as the chatbot's reply. Generative models instead use the thought vector as the basis to generate new sentences word by word using the probabilities of a word appearing given the learnt probabilities in the general language and the thought vector that was created while reading the input sentence. Generative models are much more flexible than retrieval-based models as they can generate completely new sentences while retrieval-based models give more control to the owner of the chatbot as they can decide precisely what will be said and what will not. Hybrid models also exist combining both of these aspects. For example models that attempt to generate sentences but if they fail to do so fallback to retrieval methods.

The tools used in these models are based on machine translation using encoder and decoder systems to predict the next words in a sequence. Instead of translating a sentence from one language to another, they instead "translate" an input sentence into a reply. Indeed, much work has been done already in the field of machine translation and the tools used in this domain seem to give generally decent results when used to generate replies in conversations instead, though they are certainly not at the human-level yet and are generally unsatisfying [9].

III. Evaluating chatbots

Evaluating the quality of chatbots remains a rather controversial topic has no standard metric fits the following

TABLE I
EXAMPLE IN WHICH THE AUTOMATED BLEU OR METEOR METRICS WOULD GIVE FALSE-NEGATIVES.

Context of the conversation
Speaker A: Did you hear that the new Batman movie is coming soon? Speaker B: Actually I did not. Do you think it's going to be good?
Ground truth
Speaker A: I'm not sure. They keep making new ones but they keep getting worse.
Chatbot
Speaker A: Yeah, I can't wait for it. I'm going to get tickets as soon as possible.

three criteria: automation, similarity to human judgment and precision. In this section we will review some of the methods that can be used today to evaluate conversational agents.

The TT remains a gold standard. Indeed, most users want to feel like they are conversing as easily with the chatbot as they would with a human [10]. In this case users are usually asked to evaluate how human-like the conversation felt. This method has shortcomings when it comes to being automated and does not have a good precision if no additional measures are not added. Indeed, while getting a high evaluation on the human-like aspect is the end goal, only asking the user once at the end of the conversation does not give a good indication on when during the conversation mistakes were made.

Other automated measures are often used: the task completion rate (TCR) which is especially valuable for goal-oriented chatbots which try to help users with a specific task, but cannot be applied to general purpose or chit-chat bots; the duration of the conversation can also be used, with the assumption that longer conversations mean more engagement and thus a more pleasant experience; the number of turns during the conversation, which gives another idea of the engagement and interest of the user in conversing with the chatbot. These are easy to measure but are not well correlated with the results of the TT. They also do not give insights in what went wrong when these measures give low numbers as they only inform on the general conversation rather than specific replies of the conversational agent.

Some measures instead give more specific information regarding the different turns in the conversation themselves rather than a global rating of the conversation. The most commonly used techniques are machine translation techniques like BLEU [11] and METEOR [12]. They assess how similar the generated replies are to an answer which would have been given by a human to the same question. These methods have the great advantage of being easy to automate, but have the disadvantage of not taking into account prior elements of the conversation. Besides, comparing the words being used can create false-negatives, as a perfectly intelligible and human-like response might go in an unexpected way that would be different to the sentences it would be compared to, and thus give a low score despite being perfectly valid (see an example in Table I) [13].

Artificial intelligence can also be used to evaluate the quality of chatbots. For example, RNN can be trained to mimic the evaluation of chatbots made by humans [14]. Scores given by the neural network were then significantly correlated to those given by humans on a scale of appropriateness, which the authors indicate to be the most consistent metric between human judges. Unfortunately the accuracy of such evaluation models tend to also depend on the context of the conversations (surely one would also appreciate the irony of having a chatbot emulating a human being evaluated by a similarly produced artificial judge emulating the evaluation of a human judge. It still remains an interesting first pass of evaluation). It is also possible to aggregate different metrics using trained models to emulate human judges rather than focusing on a single metric, such as engagement (captured with the number of turns or the median duration of the conversations), coherence, conversational depth, topical diversity and domain coverage [15]. The main issue here is that some ratings can be quite subjective and give a high variability. For example, the authors indicate that "A user might give a conversation 5 stars because he/she thought the socialbot was humorous, while another user might find it unknowledgeable". Thus it might be unfair to chatbot to expect them to be generally better at everything than other chatbots, while just like humans, some chatbots might be better suited than others to some tasks and not perform as well in others while remaining above an acceptable baseline.

Finally, an ideal metric would also include a rating of emotional aspects of the chatbot. Especially in conversations related to physical or mental health, having a robot show emotional skills such as empathy is an important aspect to improve how the users view and interact with the chatbot [16], [17]. These social skills would also likely be important to evaluate in the contexts of education and customer services.

Using human judges still remains the gold standard as ultimately these tools are meant to be interacting with humans. Despite the important part of subjectivity in human evaluations due to their individual expectations of a conversational partner, not all aspects of a normal human conversation are currently being encompassed by automated measures, and thus humans need to remain a part of the testing loop and the TT still has good days ahead of it before it can be fully replaced.

IV. The Turing Test as a test of humanness

While the TT was initially proposed to be a test of the intelligence of a machine [1], it has clearly shifted to being viewed as a test of humanness and is now used as such in the existing panel of evaluating metrics.

The TT, in its modern understanding of it, consists in having a human judge chat, through a text interface, with two other agents: a human and a machine. The goal for the human judge is to find which of the agent is the human, and which is the machine (or in some versions whether there is a machine at all). If after a five minute conversation the judge fails to identify the machine correctly in 50% of the trials then the machine is so much like a human that, according to Turing, it would be necessary to attribute thoughts to it in the same way we do so with humans: assuming they have mental states the way we do because they behave the way we do.

Lassègue [18] indicates that there is also another entity which is important to consider in the Turing Test: what he calls the umpire, the experimenter or the arbiter who will stop the test after a specific amount of time and tell if the judge was right or wrong, who already knows the answer to the test. This is important because the amount of time required to pass a TT varies greatly, sometimes without much justification. Turing suggests 5 minutes, but why not 10? 7? 10 minutes and 30 seconds? Past that time the chatbot could potentially reveal itself quite clearly.

The TT historically received many critics when it was suggested to be testing intelligence. One of the most famous being Searle's Chinese room argument [19]. In summary, Searle is in a room in which he is given Chinese symbols that he must reply to with Chinese symbols, along with some instructions in English (called a program) to link one input list of symbol to one output list of symbols. Unable to understand Chinese himself, Searle claims that if he was able to fool Chinese people simply by following the instructions (program) given to him in making them believe that he was Chinese himself, he still would not understand Chinese at all, and would be mindlessly following these instructions.

It is important to point out here, that Searle only applied his objection to a specific kind of AI: formal AI, using formal rules to interact through text with the user, he did not say that machines would never be able to think, but that in order to do so we would need to understand the brain rather than abstracting its general functions without understanding how it is working. A machine able to pass the Turing Test thanks to the perfect use of the manipulation of symbols would not necessarily have a mind of its own, would not necessarily think, would not necessarily be intelligent. As others pointed out, these symbols need to be grounded in one way or another, to represent something to really mean anything, thus the need for a more sensori-motor development of AI [20] along with an understanding of how the brain works and understand objects [21]. Searle indeed explains:

> As to whether or not machines will be conscious, it is important to remember that we are machines. We are biological machines and we are conscious. I do not see any reason, in principle, why we could not build an artificial machine that was conscious, but we are unable to do that now because we do not know how the brain does it. The question, "Can you build an artificial machine that is conscious?" is just like the question "Can you build an artificial heart that pumps blood?" We know how to build artificial hearts because we know how the biological heart works. We do not know how to build an artificial brain because we do not know how the brain works. But assuming we knew how the brain worked, I see no obstacle in principle to building an artificial conscious machine. The important thing to see is that the human brain is a machine, a biological machine, and it produces consciousness by biological processes. We will not be able to do that artificially until we know how the brain does it and we can then duplicate the causal powers of the brain. Perhaps we can do it in some completely different medium as we build artificial hearts in a completely different medium from muscle tissue, but at present we do not know enough about the brain to build an artificial brain. [22][1]

A similar remark was given even earlier by Shannon & McCarthy.

> A disadvantage of the Turing definition of thinking is that it is possible, in principle, to design a machine with a complete set of arbitrarily chosen responses to all possible input stimuli.... With a suitable dictionary such a machine would surely satisfy Turing's definition but does not reflect our usual intuitive concept of thinking. [24, p. vi]

Where Searle [19] makes a great leap though is when he describes the instructions, or the program in his Chinese room example (similarly the dictionary for Shannon & McCarthy [24] or the tree of sensible replies for Block [25]). They all describe a problem of the TT which is indeed real and offer conceptual examples that would pass a TT. But are these examples doable in practice? Is such a detailed and exhaustive list of instructions possible? It is extremely unlikely [26].

Indeed conversations do not follow any set of rules as strictly as one might assume. Sure, one might start a conversation with hello and end it with goodbye, as politeness would indicate. In fact, philosophers and linguists have attempted to produce a set or rules that would explain how we converse with others, starting the field of conversational pragmatics. Grice is one of them [27]. He came up with the *Cooperation Principle* (the idea that conversational partners try to cooperate during a conversation), and with four maxims that are a direct consequence of this principle: 1) the maxim of quality focusing on the truthfulness and certainty of an information given, 2) the maxim of quantity focusing on the amount of information given (neither too little nor too much), 3) the maxim of relation suggesting that participants in a conversation try to remain relevant and 4) the maxim of manner focusing on how the information is given (briefly, clearly, orderly and without ambiguity). Yet again these are not rules, they are more like expectations that each agent in a conversation has of the production of the other agents. Speakers very often do not follow them strictly: "He's a shark" is obviously not a statement that must be taken literally, but it instead conveys the idea that "he" will take everything he can from you. Grice was well aware of that and considered this practice in conversations

[1] Note that Turing himself did not seem to be against that idea. He indeed claims multiple times that in order to pass the Turing Test the best strategy would be to learn from the way humans think, though he does not dismiss the possibility that other strategies could work as well (see [23, p. 472]). The main difference between Turing and Searle being that Turing suggests this can be done at the software level while Searle considers it to only be possible at the hardware level.

to be “opting-out” of the maxims: a deviation from the maxims still within the context of a cooperation. And then there are actual violations of the maxims in the cases where participants in a conversation no longer try to cooperate: for example lying in a conversation would be a violation of the maxim of quality regarding to the truthfulness of the information given, which would be done without the knowledge of the other conversational partner: thus voluntarily removing oneself from the act of cooperation in the conversation. Still, violating these maxims does not make it less human, but any violation that is detected by the other partners will give rise to different inferences, and the violation itself will be considered to be a piece of information in its own right. The most important concepts that Grice offers which is of importance for chatbots is the distinction between what is *said* and what is *meant*. Take the following example: “Come in! But I do not have alcohol”. At face value, it would be difficult to tell how inviting someone in would be this directly related to alcohol without any other information. Yet this sentence is easily understood and can trigger an offended reply, a disappointed one or an amused one, depending on the relationship between the two participants in the conversation. What is *meant* here is “Come inside, but there is no alcohol inside and I know you might have expected that we would share alcohol”. Here the key to understand the mention of alcohol (drinking was expected) is completely implicit. While the mention of the alcohol seem to be coming out of nowhere and thus violating the maxim of relation, it is understood as being perfectly relevant within the given context, because of prior expectations about the situation.

To explain such productions, Sperber and Wilson developed the *Relevance Theory* [28]. The main idea behind it is that participants in a conversation actively search for relevance in the utterances of others. The Relevance Theory describes an utterance with optimal relevance as an utterance which has the greatest contextual effect on the listener’s mental representations for the least cognitive cost (least effort in retrieving what is *meant* from what is *said*). Indeed, in the previous example what is the use of including that drinking is something expected when it’s an expectation both participants in the conversation already share? This would only be making the interpretation of the sentence harder, would take more time, and would not add anything (it would not change the mental representations of the listener as this is something they would already know). Thus adding it in the utterance is irrelevant, and it remains implicit.

Because the Relevance Theory expects participants in a conversation to have an idea of what is in the mind of the others, participating in an actual conversation (at the human level) requires a Theory of Mind [29]. The Theory of Mind is the concept according to which humans (among others animals) are mindreaders. Not in the metaphysical way of course, but humans understand that other humans also think, that they have mental representations of the world that might be different or similar to their owns. There are things others do not know that we know, and there are things we do not know that they might know... It is the reason why there are questions in conversations: we understand that others might have the answers we are looking for, and we ask them to share the information they have with us. Reciprocally, the only reason why people answer questions is because they assume people who ask them do not already know the answers and will learn (their mental representations will change once the answer is given to them). Evidence indicate that humans acquire this capacity very early in their life [30], [31] and the presence or absence of this ability in other species is still a strong debate in the scientific community [32], [33], which is not entirely surprising given the difficulty in finding ways to explicitly communicate the question without ambiguity to young humans [30].

Because conversations are built around these principles, the replies given during a conversation are not fixed and will depend heavily on what each participant in a conversation believes the other knows. Thus, predefined rules, as mentioned in Searle’s Chinese room (the instructions given to the man inside the room on how to match a string of symbols as a reply to another string of symbols) cannot do more than imitate and drastically reduce the range of possibilities that natural conversations have. Not only would the set of instructions be infinitely large, but it would also need constant updating to be tuned for the specific audience and for the changes in time as the natural language evolves. Thus, it is our belief that rule based AI such as ELIZA, A.L.I.C.E, and Kuki.ai will not be able to reliably pass the TT for their inability to learn from their interactions. Similarly, retrieval based systems using machine-learning in order to detect the intent still likely will not be able to reliably pass the TT as they are not able to generate new answers that would fit new situations like a human would. Only a machine learning to infer meaning and to change how it expresses itself should be able to pass the Turing Test reliably, even though currently generative AIs are less useful and more frustrating than retrieval-based AIs.

But would a judge be able to tell the difference in a TT? Is not understanding context and the mind of others enough to significantly prevent a machine from passing the TT? As we will show in the next section, the answer is yes.

V. Understanding humans

Comparing how human-chatbot conversations with human-human conversations has many benefits for the two fields of psychology and computer science. Investigating how humans behave compared to chatbots can help us make better chatbots, and investigating interactions with chatbots can give us valuable information on what humans expect of a conversational partner. And yet, despite the fact that the TT can be used as en experimental paradigm useful to explore human expectations in conversations, it is remarkably absent in international publications in the field of psychology and pragmatics. Indeed, doing a quick search on Google Scholar reveals about 33.000 entries for “Turing Test” while adding keywords from the field of pragmatics makes the search drop to below 300 entries (“Turing Test” Implicatures: 209 results,

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

"Turing Test" "Relevance Theory": 96 results, "Turing Test" "Cooperative Principle": 104 results), most only mentioning these topics without focusing on them.

Chatbots are still quite far from meeting human expectations of a conversational partners. surveys and studies showing that people get quickly frustrated when using them are not hard to find (see [10], [34], [35] to name only a few). An extensive survey conducted on the literature of chatbots indicates many of the current challenges they still face [36], especially regarding social characteristics of the chatbots. This feeling of frustration can be mitigated when it is made clear to the user what can be expected of the chatbot. For example, Woebot clearly sets its users' expectations beforehand which allows the users to adapt their own behavior [37]. In the case of this chatbot (which acts as a coach to help deal with anxiety and depression), the bot remains in control of the conversation at all times as the user navigates pre-defined decision trees, and in doing so it is able to carry out its task, though in cases that are too severe the user is redirected to a hotline through which they can interact with professionals to seek help. This transparency about the chatbot's abilities (along with its very sparse use of natural language understanding) allows it to be efficient in its task of helping people cope with anxiety and depression, at least for a short time (as the study did not investigate long term effects). Similar effectiveness of this chatbot seems to be observed to help control substance use [38].

Indeed, the closer the chatbot is to feeling like a human, the more users will be expecting human-like abilities in their interactions with them. It is possible to observe this effect even on the same chatbots depending on how they are introduced. For example one study can find the bot entertaining enough for the users to keep conversing with it for extended conversations despite a quality of conversations significantly lower than that with humans [39], while another can observe judges in a TT being quite perplexed when they are not made aware that the author of the messages might be a chatbot, wondering whether the person writing such messages might be "mentally ill" [40].

These situations of violating the user's expectations are common when interacting with chatbots, creating a feeling similar to Mashiro Mori's uncanny valley [41] which is a famous effect observed with robots [42] (The closest an artificial agent, robot or chatbot, gets to human behavior or appearance, the greater the expectations of humans interacting with it will be, and the greater the frustration or uncomfort if they are not meant). Still one might wonder if all other aspects remaining similar, violating such expectations would be enough to prevent a machine from passing the TT. Saygin & Cicekli [40] investigate this issue by trying to assess to what level each of Grice's maxims [27] has (or does not have) an effect on the participants responses in a TT. Their findings indicate that not all the maxims have similar influences on the answers in the TT. Indeed, violations of the maxim of manner (which deals with how information is given to the user) has no detrimental effect on the judges' perception of the humanness of the chatbot. In fact, they even observe that it has a positive effect as long as no other maxims are violated. They explain this finding by the fact that violating this maxim can produce a seemingly more emotionally loaded reply, emotions being a feature more readily (and understandably) associated to humans than to machines. Violations of the maxim of quantity was shown to have either no effect on the TT (when the maxim was violated to give too little information) or to be quite detrimental to perceived humanness (when the maxim was violated to give too much information, giving an encyclopedic feel to the reply). The difficulty with assessing the individual effect of violations of this maxim is that when violated it also has the tendency of violating the maxim of relation, which produced by far the strongest adverse effect to the feeling of humanness: the judge was left feeling like the chatbot simply did not understand the question (or did not want to talk about this topic for no understandable reason when the judges were not aware that a chatbot was present). Finally, the authors were unable to show a specific influence of the maxim of quality on the humanness of the chatbot for it also had a tendency of being violated along with other maxims.

An important difference remains between the above paper and a regular TT. In Saygin & Cicelki's paper judges were reading excerpts of conversations recorded during a Loebner prize competition and did not actually interact with the chatbots. Would users interacting with a chatbot for which the only issue would be a lack of relevance or other violations notice this flaw enough to correctly label the chatbot as a machine? We have tried to answer this question in previous papers [43]–[45] by inviting participants to play the judges in a TT. The main interest in our approach here was to test the influence of these violations only: indeed the judges participated in two conversations in a random order, being informed that one would be with a chatbot and the other with a human. In truth there was no chatbot at all. Indeed using one would have made it more difficult to test whether or not the observed differences would have been caused by the violations or by other factors related to the chatbot. Both conversations were played by the same human experimenter, each time portraying a fictive character (the same fictive character between the two conversations), except that in one conversation the experimenter was tasked to produce violations of one of Grice's maxims. Once again, the violations which had the most effect on the feeling of humanness were violations of the maxim of relation [43], [45] and violations of the maxim of quantity giving rise to an encyclopedic feeling [44]. This effect was also visible in the delay between the experimenter's utterance and the participant's turn (which is longer following a violation than following an expected reply), further indicating that these violations are indeed the cause of the observed difference. In addition, the kind of violations of the maxim of relation in these papers were slightly more subtle than the blatant violations that can often be found in chatbots: the experimenter was not allowed to use previous knowledge of the conversation in their replies, but could still answer relevantly if all the necessary information to do so was contained in the participant's last message. For example:

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Human: Do you like reading?
Experimenter: Not really no. It's not really my thing.
Human: Why not?
Experimenter: It's hard to tell. Do you have any brothers or sisters?

In the experimenter's first reply they are allowed to give a relevant answer, but in their second answer they were not allowed to use the knowledge that the topic was about reading. Thus they used a generic reply instead, producing a violation of the maxim of relation.

This type of violation is very easy to get on any chatbot currently available. Asking generic questions such as "Why?" or "Why not?" requires the chatbot to use the context of the message (the conversation's history) to be able to reply correctly. In the human's second question here, they assume that their reader still has in mind the topic of the conversation (reading is not the experimenter's thing), while the experimenter must infer that what the participant *means* is "Why is reading not really your thing?" when they *say* "Why not?".

More studies need to be carried out to explore just how sensitive the TT is to even more subtle violations, but with the evidence at our disposal today, it seems highly likely that only a chatbot able to converse in a relevant way in every situation would be able to pass the TT (especially in its 3 players version: judge, machine and human, with no limits on the topics of discussion), and this would require the ability to develop an idea of what is relevant to the user, and thus for the chatbot to have a theory of mind [46]. We are not there yet [4].

VI. Conclusion

While we have only scratched the surface of the literature regarding the TT, we explored the existing literature discussing the importance of conversational pragmatics within chatbots, and we have attempted to show how the TT is a very relevant tool in evaluating the ability of chatbots to generate relevant replies in an open conversation which is (so far) not matched by any other evaluation method.

We have also discussed how the TT in its design suggests that only an agent with a theory of mind could reliably pass it, though of course it does not set any requirements on how this theory of mind is implemented.

We also believe that the TT should be more widely used in human sciences like psychology, especially in the case of studying reasoning and conversational pragmatics. It is still a tool that is extremely rarely used despite being a valuable experimental paradigm which enables experimenters to collect direct measures (the response in the TT) and indirect measures (the delay between utterances during the conversations for example). This area of research is still underdevelopped despite its great potential for fundamental and applied research. One example is to test the influence of the use of textisms (SMS language) on the cognitive cost of processing messages in a conversation [47], or to use chatbots to investigate how behaviors are influenced by different pragmatic clues in the ultimatum game [48].

Finally, some readers might object that we did not settle the issue of whether passing the TT proves that one has a mind. After all, do we need a mind to have a theory of mind?

References

[1] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, pp. 433–460, 1950.

[2] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[3] R. S. Wallace, "The anatomy of alice," in *Parsing the Turing Test*. Springer, 2009, pp. 181–210.

[4] B. Jacquet and J. Baratgin, "Mind-reading chatbots: We are not there yet," in *International Conference on Human Interaction and Emerging Technologies*. Springer, 2020, pp. 266–271.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] M. Qiu, F.-L. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and W. Chu, "Alime chat: A sequence to sequence and rerank based chatbot engine," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 498–503.

[7] M. Dhyani and R. Kumar, "An intelligent chatbot using deep learning with bidirectional RNN and attention model," *Materials Today: Proceedings*, vol. 34, pp. 817–824, 2021.

[8] X. Qiu and X. Huang, "Convolutional neural tensor network architecture for community-based question answering," in *Twenty-Fourth international joint conference on artificial intelligence*, 2015.

[9] B. Wei, S. Lu, L. Mou, H. Zhou, P. Poupart, G. Li, and Z. Jin, "Why do neural dialog systems generate short and meaningless replies? a comparison between dialog and translation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7290–7294.

[10] M. Jain, P. Kumar, R. Kota, and S. N. Patel, "Evaluating and informing the design of chatbots," in *Proceedings of the 2018 Designing Interactive Systems Conference*, 2018, pp. 895–906.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[12] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[13] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," *arXiv preprint arXiv:1603.08023*, 2016.

[14] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau, "Towards an automatic turing test: Learning to evaluate dialogue responses," *arXiv preprint arXiv:1708.07149*, 2017.

[15] A. Venkatesh, C. Khatri, A. Ram, F. Guo, R. Gabriel, A. Nagar, R. Prasad, M. Cheng, B. Hedayatnia, A. Metallinou *et al.*, "On evaluating and comparing open domain dialog systems," *arXiv preprint arXiv:1801.03625*, 2018.

[16] M. de Gennaro, E. G. Krumhuber, and G. Lucas, "Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood," *Frontiers in Psychology*, vol. 10, p. 3061, 2020.

[17] S. Devaram, "Empathic chatbot: Emotional intelligence for empathic chatbot: Emotional intelligence for mental health well-being," *arXiv preprint arXiv:2012.09130*, 2020.

[18] J. Lassègue, "What kind of turing test did turing have in mind?" *Tekhnema: Journal of Philosophy and Technology*, vol. 3, pp. 37–58, 1996.

[19] J. R. Searle *et al.*, "Minds, brains, and programs," *The Turing Test: Verbal Behaviour as the Hallmark of Intelligence*, pp. 201–224, 1980.

[20] S. Harnad, "What's wrong and right about searle's chinese room argument?" 2001. [Online]. Available: http://cogprints.org/4023/

[21] J. Hawkins, M. Lewis, M. Klukas, S. Purdy, and S. Ahmad, "A framework for intelligence and cortical function based on grid cells in the neocortex," *Frontiers in Neural Circuits*, vol. 12, p. 121, 2019.

[22] D. Turello, “Brain, mind, and consciousness: A conversation with philosopher john searle,” 2015. [Online]. Available: https://blogs.loc.gov/kluge/2015/03/conversation-with-john-searle/

[23] A. P. Saygin, I. Cicekli, and V. Akman, “Turing test: 50 years later,” *Minds and machines*, vol. 10, no. 4, pp. 463–518, 2000.

[24] C. E. Shannon, J. McCarthy *et al.*, *Automata studies*. Princeton University Press Princeton, NJ, 1956, vol. 11.

[25] N. Block, “Psychologism and behaviorism,” *The Philosophical Review*, vol. 90, no. 1, pp. 5–43, 1981.

[26] D. McDermott, “On the claim that a table-lookup program could pass the turing test,” *Minds and Machines*, vol. 24, no. 2, pp. 143–188, 2014.

[27] H. P. Grice, “Logic and conversation,” in *Syntax and Semantics 3: Speech arts*, P. Cole and J. L. Morgan, Eds. New-York: Academic Press, 1975, pp. 41–58.

[28] D. Wilson and D. Sperber, *Relevance Theory*. Oxford: Blackwell, 2004, pp. 607–632.

[29] D. Premack and G. Woodruff, “Does the chimpanzee have a theory of mind?” *Behavioral and brain sciences*, vol. 1, no. 4, pp. 515–526, 1978.

[30] J. Baratgin, M. Dubois-Sage, B. Jacquet, J.-L. Stilgenbauer, and F. Jamet, “Pragmatics in the false-belief task: let the robot ask the question!” *Frontiers in Psychology*, vol. 11, p. 3234, 2020.

[31] I. Bretherton, S. McNew, and M. Beeghly-Smith, “Early person knowledge as expressed in gestural and verbal communication: When do infants acquire a “theory of mind”,” *Infant social cognition*, vol. 333, p. 73, 1981.

[32] C. Krupenye and J. Call, “Theory of mind in animals: Current and future directions,” *WIREs Cognitive Science*, vol. 10, no. 6, p. e1503, 2019.

[33] D. C. Penn and D. J. Povinelli, “On the lack of evidence that non-human animals possess anything remotely resembling a ‘theory of mind’,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1480, pp. 731–744, 2007.

[34] P. B. Brandtzaeg and A. Følstad, “Chatbots: changing user needs and motivations,” *Interactions*, vol. 25, no. 5, pp. 38–43, 2018.

[35] E. Luger and A. Sellen, “” like having a really bad pa” the gulf between user expectation and experience of conversational agents,” in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 5286–5297.

[36] A. P. Chaves and M. A. Gerosa, “How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design,” *International Journal of Human–Computer Interaction*, pp. 1–30, 2020.

[37] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial,” *JMIR Ment Health*, vol. 4, no. 2, p. e19, June 2017.

[38] J. J. Prochaska, E. A. Vogel, A. Chieng, M. Kendra, M. Baiocchi, S. Pajarito, and A. Robinson, “A therapeutic relational agent for reducing problematic substance use (woebot): Development and usability study,” *J Med Internet Res*, vol. 23, no. 3, p. e24850, March 2021.

[39] J. Hill, W. Randolph Ford, and I. G. Farreras, “Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations,” *Computers in Human Behavior*, vol. 49, pp. 245–250, 2015.

[40] A. P. Saygin and I. Cicekli, “Pragmatics in human-computer conversations,” *Journal of Pragmatics*, vol. 34, pp. 227–258, 2002.

[41] J. Vallverdú, H. Shah, and D. Casacuberta, “Chatterbox challenge as a test-bed for synthetic emotions,” in *Creating Synthetic Emotions through Technological and Robotic Advancements*. IGI Global, 2012, pp. 118–144.

[42] C. F. DiSalvo, F. Gemperle, J. Forlizzi, and S. Kiesler, “All robots are not created equal: the design and perception of humanoid robot heads,” in *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, 2002, pp. 321–326.

[43] B. Jacquet, J. Baratgin, and F. Jamet, “Cooperation in online conversations: the response times as a window into the cognition of language processing,” *Frontiers in psychology*, vol. 10, p. 727, 2019.

[44] B. Jacquet, A. Hullin, J. Baratgin, and F. Jamet, “The impact of the gricean maxims of quality, quantity and manner in chatbots,” in *2019 international conference on information and digital technologies (idt)*. IEEE, 2019, pp. 180–189.

[45] B. Jacquet, J. Baratgin, and F. Jamet, “The gricean maxims of quantity and of relation in the turing test,” in *2018 11th international conference on human system interaction (hsi)*. IEEE, 2018, pp. 332–338.

[46] B. Jacquet and J. Baratgin, “Towards a pragmatic model of an artificial conversational partner: opening the blackbox,” in *International Conference on Information Systems Architecture and Technology*. Springer, 2019, pp. 169–178.

[47] B. Jacquet, C. Jaraud, F. Jamet, S. Guéraud, and J. Baratgin, “Contextual information helps understand messages written with textisms,” *Applied Sciences*, vol. 11, no. 11, 2021.

[48] B. Beaunay, B. Jacquet, and J. Baratgin, “A selfish chatbot still does not win in the ultimatum game. [accepted],” in *6th International Conference on Human Interaction & Emerging Technologies: Future Systems*. Cham: Springer International Publishing, October 2021.