Psychological Topics, 31 (2022), 1, 185-202
Original Scientific Paper UDC: 159.9.072:004.8
https://doi.org/10.31820/pt.31.1.9

Would You Pass the Turing Test? Influencing Factors of the Turing Decision

Adrienn Ujhelyi, Flora Almosdi, and Alexandra Fodor
Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary

Abstract

We aimed to contribute to the emerging field of human-computer interaction by revealing some of the cues we use to distinguish humans from machines. Maybe the most well-known method of inquiry in artificial intelligence is the Turing test, in which participants have to judge whether their conversation partner is either a machine or human. In two studies, we used the Turing test as an opportunity to reveal the factors influencing Turing decisions. In our first study, we created a situation similar to a Turing test: a written, online conversation and we hypothesized that if the other entity expresses a view different from ours, we might think that they are a member of another group, in this case, the group of machines. We measured the attitude of the participants ($N = 100$) before the conversation, then we compared the attitude difference of the partners to their Turing decision. Our results showed a significant relationship between the Turing decision and the attitude difference of the conversation partners. The more difference between attitudes correlated with a more likely decision of the other being a machine. With our second study, we wanted to widen the range of variables and we also wanted to measure their effect in a more controlled, systematic way. In this case, our participants ($N = 632$) were exposed to an excerpt of a manipulated Turing test transcription. The dialogues were modified based on 8 variables: humour, grammar, activity, the similarity of attitude, coherence, leading the conversation, emoji use, and the appearance of the interface. Our results showed that logical answers, proper grammar, and similar attitudes predicted the Turing decisions best. We also found that more people considered mistaking a computer for a human being a bigger problem than vice versa and this choice was greatly influenced by the participants' negative attitudes towards robots. Besides contributing to our understanding of our attitude toward machines, our study has also shed light on the consequences of dehumanization.

Keywords: Turing test, artificial intelligence, attitude, social psychology

Introduction

Artificial intelligence is one of the fastest developing areas of technology, and it is essential to understand the underlying mechanism of human-computer

✉ Adrienn Ujhelyi, Institute of Psychology, ELTE Eötvös Loránd University, Izabella u. 46, Budapest 1064, Hungary. E-mail: *ujhelyi.adrienn@ppk.elte.hu*

interactions as they most probably play an important role in humanity's future. In the – maybe not so distant – future we will need to acquire a new skill: be able to distinguish human beings from machines to maintain our feeling of understanding and controlling our world. Either mistaking a machine for a human being (anthropomorphization) or mistaking a human being for a machine (dehumanization) can hold serious consequences.

The classic paradigm that was based on the abovementioned human skill was the Turing test (Turing, 1950), though its original aim was quite different, it was rather designed to measure the intelligence of machines. The test created a situation where a human judge had to carry on two written conversations: one with a human partner, and another with a computer program. After the conversations, they had to decide which one of the conversational partners was the human and which was the computer program. Turing theorized that a machine could be regarded as intelligent (meaning passing the test) if it could trick interrogators to believe it was a human (Turing, 1950).

The Turing test has been considered a milestone in the research of artificial intelligence and has been widely criticized at the same time since its introduction (Block, 1981; Halpern, 2006; Harnad, 1991; Saygin et al., 2000; Searle, 1980; Weizenbaum, 1976). Alan Turing himself argued that the question 'Can machines think?' could be misinterpreted easily (Turing, 1950), and instead we should ask another one: 'Is the Imitation Game the right way to measure the intelligence of machines?' One of the main concerns was that intelligence is way more than comprehensible talk, and a computer can use a bunch of symbols and give the right answers without knowing what they are doing (Searle, 1980). In addition to this, it is still questionable whether there is such a thing as "intelligence in general" (French, 1990). In the Turing test, intelligence is identified as human intelligence and the machine is labelled to be intelligent if it could mislead the human interrogator by imitating human behaviour. Hayes and Ford (1995) even suggested that the fact that the interrogators of the Turing test are humans, could be the "Achilles heel of the test" - it is something that makes its reliability a lot lower. Collins (1990) even though the test is about the observer, not about the chatbot or the machine.

Social Aspects of the Turing Test

If we look at the Turing situation from another perspective, the behaviour of the judge is even more interesting than the intelligence of the machine. By learning why some participants ascribe intelligence to the chatbots, making the test falsely positive, or why others have mistaken the human for a chatbot, we will have a deeper understanding of the Turing test and ultimately what constitutes humanness.

Some studies tried to reveal the most influential factors of the Turing decision. Lortie and Guitton (2011) for example conducted linguistic analysis on the transcripts describing existing Turing dialogues where the human subjects have been

judged as robots. They analysed the descriptive and cognitive parameters of the conversations. According to their results, people judged as robots used fewer words, fewer articles, and fewer compliments per post than those judged as humans.

The causes of misidentification were the focus of the Warwick and Shah's study (2015). They conducted content analyses on transcripts of the Turing test and found that the most important factors when misjudging humans were the following: lack of shared knowledge, out-of-the-box answers, boring answers, dominating the conversation.

Candello et al. (2017) demonstrated that not just the content of the conversation but even visual aspects, such as typefaces can influence the perception of humanness. Machine-like typeface (OCR-A) biased participants towards perceiving the entity as a machine but, unexpectedly, a handwritten-like typeface (Bradley) did not have the opposite effect. Those effects were influenced by the familiarity of the user with artificial intelligence and other participants' characteristics.

Besides these studied variables other psychologically important factors may influence the Turing decision. The literature on the social psychology of intergroup relations can offer some clues. For example, attitude difference is a consensually influential factor when categorizing someone as an ingroup or an outgroup member (Balliet et al., 2014; Efferson et al., 2008; Tajfel, 1969). We tend to rate those with whom we belong to the same group more positively than outgroup members (Tajfel, 1969), but if this person has negative or stigmatized characteristics, we usually distance them from ourselves and even reject them, because they endanger our sense of positive identity, thus, we deny them the membership to our group (Novak & Lerner, 1968).

In sum, we can state that the Turing decision is not an objective, rational one, it is influenced by many psychological factors on the judge's side. With our studies, we wanted to explore some of the potentially influential factors.

Study 1

Objective of the Study

Previous studies focused on the impact of conversational features on the Turing decision (e.g. Candello et al., 2017; Lortie & Guitton, 2011; Warwick & Shah, 2015), other, namely social psychological factors are still understudied in this field. For our first study, we aimed to examine the effect of attitude-differences on the Turing decision, as there is no other study on this relationship yet. We hypothesized that during an interaction where the group-membership (that is if they are human beings or computer programs) of interaction-partners is unknown to participants, perceived attitude-difference will affect the decision about the partner. Namely, if they perceive attitude differences, they will be more likely to judge their partner a computer

program (ergo an outgroup member). We also wanted the participants to reflect on their decision, so we also asked them to explain their decision in order to identify the main categories.

Method

Participants

One hundred undergraduates from the Eötvös Loránd University were recruited from a university pool and reached via e-mail. Their mean age was 21.49 years (*SD* = 2.32), 74% of them were women. As there is no Hungarian speaking chatbot available today, we conducted the whole study in English – every test was administered in English, not just the conversation part. Thus, the requirement for participation was an advanced level of English, but most of our participants were not native speakers (87%).

Measures and Procedure

For the research, we developed a software application (see the Appendix) that was used for both the chat and administering the psychological tests. The application itself contained four separate parts, namely a Test Code Generator (TCG) that generated random user identifiers (UIDs), a Client-Side Application (CSA) running on the participants' browsers, a Server-Side Application (SSA) managing participants' sessions and keeping track of their results and finally a Database Application (DBA) to store the data collected from the participants' actions. For details, see Appendix.

When participants agreed to take part in our research, they were contacted via email. Participants' self-reported knowledge of the English language on a scale from 1 (*able to have a simple conversation*) to 10 (*can communicate fluently*) was 7.3 (*SD* = 1.5).

They agreed to be present and ready in front of their computer, with a good internet connection at a previously appointed time. At any given time, the minimum number of participants was 2, the maximum was 20. We used a Viva Voce Turing test, meaning that there were only two conversational partners at the same time, as opposed to the original setup where the judges had to simultaneously compare two hidden interlocutors (Warwick & Shah, 2014). Participants received a code that was randomly generated before the study. This code was their password for the Turing research software and their identification number as well. This solution secured the anonymity of our participants and allowed us to match the conversational partners' data during our analysis without using any of their personal information.

After reading and accepting the form of Informed Consent, participants had to type their code into the research software, which forwarded them to the attitude-scale phase.

1. *Measuring attitudes.* As we wanted to measure real attitudes and create real-life differences in attitudes, we had chosen attitudes towards a topic that was most likely to generate an argument among people. Based on a pilot study ($N = 42$, mean age 26.71 (18-52), 42% women) we have chosen the *Attitudes Towards Prostitutes and Prostitution Scale* (ATPP; Levin & Peled, 2011) as it proved to have the best reliability value (Cronbach $\alpha = .81$). So, the participants of the main study had to answer the 29 items of the ATPP Scale, rating their level of agreement on a 7-point Likert-type scale (1 - *strongly disagree*, 7 - *strongly agree*). The reliability of the ATPP scale after deleting the lowest scored items on item-total correlations (6 items) was acceptable (Cronbach $\alpha = .77$).
2. *Conversation with an entity.* Instruction to the conversation phase stated that they were going to talk with an entity that could be either a human being or a chatbot, and they had to talk about a given topic (”According to your opinion, prostitution should be legalized, or not? Why?”), and they were warned not to share any personal information about themselves, for the sake of protecting their anonymity. We restricted the number of characters sent at the same time (150 characters), and the duration of the conversation (10 minutes), to create a dynamically flowing conversation.
3. *Turing decision*. After 10 minutes, the software automatically forwarded them to the Turing decision phase, where they had to indicate whether their partner was a human being or a chatbot, and they were also asked to explain their decision in an open-ended question. At this point, we must state clearly, that as we were only interested in the human aspect of the Turing-type test, all of the conversational partners were human participants, randomly paired by the software based on their codes.
4. *Additional information*. Finally, participants had to answer questions related to their demographics (gender, age, localization, education, and knowledge of the English language).
5. *Debriefing*. In the end, they entered the debriefing phase, where they could read a short text explaining the aim of the study.

Our research was approved by the Ethics Committee of Eötvös Loránd University.

Results

The results of the Turing decision revealed that 42% of participants ($N = 42$) thought that their conversational partner was a chatbot.

The indicator of attitude-difference was calculated by taking the attitude-points of each conversational partner and subtracting one from the other. Thus, we got the distance (0 to 43, mean = 27) between the two conversational partners. After checking the assumptions (a dichotomous dependent variable with mutually exclusive categories and Box-Tidwell test for linearity), we performed a logistic regression to ascertain the effect of attitude difference on the Turing decision. The logistic regression model was statistically significant, $\chi^2(1\ N = 100) = 23.402$, $p < .043$. The model explained 12.0% (Nagelkerke R^2) of the variance, meaning that those with more different attitudes from each other were more willing to categorize their partner as a robot.

We have also analysed the explanations the participants gave for their decision (Table 1). Out of the 100 participants, 98 wrote an explanation (out of them 31 answered: ”I don’t know.” First, we created 2 categories based on their decision (the partner was a human or a robot) and categorized them separately. The most frequent categories for the Human group were: (1) grammar and style mistakes (mentioned 15 times); (2) being interactive, giving adequate answers (5 mentions); (3) some of the answers lacked argument, mentioning just a feeling (5 times); (4) not answering immediately (3 times); (5) showing emotions or empathy (3 times). For the Robot group, the categories were very similar, only their order of frequency was different: (1) not being interactive enough, giving out of context or too generalized answers (14 times); (2); being slow in answering (10 times); (3) writing grammatically precisely (8 times); (4) showing lack of emotions (5 times). We found no difference between the two groups of participants regarding gender, age, or knowledge of the English language.

Table 1

Categorization of the Explanations in Study 1

Categories	Decision: human ($N = 58$)		Decision: robot ($N = 42$)	
	frequencies	meaning, example	frequencies	meaning, example
grammar	15	bad grammar: ”I think it was a human, because of the many grammatical mistakes, the misuse of words, the missing words and so on.”	8	too good grammar: ”The sentences seemed too perfect to me, no typos, always writing \”I am\” instead of \” I’m\” and using sophisticated words.”
interaction	5	interactive, adequate answers: ”I believe I have been talking to a human because his/her answers were reactions to my ones”	14	not interactive, out of context replies: Cthe whole conversation, it feels like I was getting either generic or random responses overall.”

Categories	Decision: human ($N = 58$)		Decision: robot ($N = 42$)	
	frequencies	meaning, example	frequencies	meaning, example
vague answer	5	just a feeling "She/he reacted to my opinion like a real human" "I just felt like I was talking to a human"	-	-
reaction time	4	slow (took their time): First of all my partner took a couple of minutes to answer which according to my experience with chatbots is very unusual.	10	slow: "I think the replies were just fine. But I think if I would've been talking to a human he/she would've answered me much quicker."
emotion	3	being emotional, showing empathy: "My partner seemed to be very friendly and open. He showed emotional support"	5	lack of emotions: "I think my partner didn't really show any emotion in his answers. I think this topic should call for some negative emotions. He was too rational for me."
I don't know or no answer	26		5	

We were particularly interested in the answers referring to attitude (using the terms attitude, opinion, or attitude difference). We have found 9 such answers (Table 2).

Table 2

Categorization of the Explanations Related to Attitude or Opinion in Study 1

	Similar attitudes	Opposing attitudes
Human	- "Because it felt real, and the person I talked to was very agreeing and talked further from my point." - "My partner and I had the same opinions and also gave some good arguments that I think only a human can come up with" - "To me, it seemed like I was talking to a human, because their responses and reactions coordinated with what I was saying, even at the beginning, before we started the actual conversation about the topic"	- "Because they had a very strong opinion, opposing to mine."

	Similar attitudes	Opposing attitudes
Robot	- ”It always reflected my opinion.” - ”He wasn’t really expressing his own views, just agreeing with mine” - ”I think it was a chatbot because it was saying what I said. It didn’t have its own opinion”.	- ”Because that other kind forced me to think like him or expect these sayings. Even though everyone has their own opinion.” - ”Because it gives the opposite answer. And I think most humans would totally agree with me.”

Discussion of Study 1

Almost half of our participants (42%) decided that their conversational partner (that was in every case a human being) was a computer program. Previous studies showed (Shah & Henry, 2005; Shah & Warwick, 2010) that while this type of misidentification (they called it confederate effect) is usually less frequent (2 out of 9 cases), it is certainly an existing phenomenon: ”The study also reveals the existence of the Confederate Effect: both female and male hidden humans in Loebner (2003) were sometimes considered machine-like from their conversation.” (Shah & Henry, 2005, p. 4). While these authors did not give an explanation, it occurred to us, that maybe the two choices have different values for the participants in the sense that one of them may seem as the riskier or more embarrassing type of error. This hypothesis is yet to be proven.

Our results also showed significant relationship between attitude-difference and the Turing-type decision, those who differed more in their attitudes were more prone to categorize their partner as a chatbot. This is in line with previous social psychological studies, for example, Schwartz and Struch’s (1989) theoretical framework that states that the perceived humanity of the outgroup largely depends upon the perceived similarity of the groups’ values. So when people see others’ values as incongruent, they are likely to perceive them to lack shared humanity.

The content analysis of the explanations revealed four main categories of reasoning: grammar, being interactive, reaction time, and showing emotions. These findings are in accordance with the literature (Candello et al., 2017; Lortie & Guitton, 2011; Warwick & Shah, 2015) showing that attributes of the text and partner such as the lengths of the conversation, grammatical mistakes, use of humour, the level of activity during the conversation play an influential role when deciding about the identity of the entity in a Turing situation.

What more can be seen is that a dimension is more salient (thus mentioned more) when the given attribute is missing (lack of emotions or not being interactive) and that the very same attribute can have very different meanings and connotations based on our perception of the partner (in the case of humans being slow was a sign of taking their time to think about the answer, but not giving an immediate answer was a telling sign of a not developed enough technology when the answer was a

robot). Some of the answers were lacking any arguments, indicating that the judge had a vague intuition about the other. Interestingly more of these answers belonged to the human group (31 for the Human group vs. 5 for the Robot).

In the next study, we wanted to include more cues and examine them more systematically. We decided that reaching this aim requires more controlled situations so instead of asking participants to engage in a real conversation we had manipulated existing conversations according to the variables sought to test and they had to rate those manipulated excerpts.

Study 2

Objectives of the Study

Our first aim with the second study was to give the topic a wider perspective by searching for the most relevant factors influencing the decisions of the judges in a situation similar to the Turing test. According to previous findings in the literature, attributes of the text and partner such as the lengths of the conversation, use of humour, the similarity of attitudes, the level of activity during the conversation play an influential role in the decision (Lortie & Guitton, 2011; Warwick & Shah, 2015). Besides, we have also used the categories from the first study, namely using proper grammar, being interactive, and showing emotions.

The second objective of the study was to test the hypothesis that came to our mind when tried to explain the results of the first study, namely to ask our participants explicitly which type of Turing decision holds a greater risk: accepting a computer program as a human being (anthropomorphization) or wrongly deciding that a human being was a computer program (dehumanization). We suspected that the participants' attitude toward artificial intelligence can play a role in answering all our questions, so we also administered the Negative Attitude Towards Robots Scale (NARS, Nomura et al., 2006).

Method

Participants

The participants were recruited from a pool of university students who earned credits for participating. The total number of participants was 632. Their ages ranged from 18 to 59 years, with an average age of 25.3 years, 289 were male (45%).

Procedure

First, the participants completed an anonymous online questionnaire containing demographic questions.

After this phase, they saw a *screenshot of a fake excerpt from a Turing test* (one example can be seen in Figure 1), a conversation between a judge and an unknown

entity. As we intentionally wanted to study specific dimensions, we manipulated the conversations based on the research of Warwick and Shah (2015) and our previous study. We used 8 independent variables, with two conditions of each: humour, grammar, activity, the similarity of attitude, coherence, leading the conversation, emoji use, and the appearance of the interface. Thus we created 16 transcripts altogether. Each participant saw only one conversation, randomly selected by the program. After reading the conversation they had to decide if the entity was human or a robot. This time we wanted to have more refined answers so instead of a nominal variable we used a scale of 1-5, with 1 being "*I am sure it's a robot*", 3 being "*I cannot decide*" and 5 "*I am sure it's a human*".

Figure 1

Example of a Manipulated Turing Type Transcript (Variable Here: the Appearance of the Interface)

>> 15:23:10 **Judge** joined in.

Judge: So I'm supposed to ask you tricky questions?

Entity: Ask me anything!

Judge: What's your fav color?

Entity: Blue, like the sky on a sunny Sunday afternoon. Can I guess yours?

Judge: Are you a dog or a cat person?

Entity: Sadly, my apartment is too small for having pets. Maybe I'll have a garden one day.

Judge: And what animal would you keep in that garden?

Entity: Giraffes are my favourites, but nothing smaller than a hippo.

Judge: You're a big dreamer, huh?

Entity: Dreaming is the only way out of a meaningless life.

>> 15:28:10 **Entity** left the conversation.

After the decision, the participants had to rate the behaviour of the entity on 8 dimensions (1 *strongly disagree* – 5 *strongly agree*) (being funny, showing similar attitudes to the judge, how realistic were their answers, whether they dominated the conversation, how logical were their answers, whether they played an active role during the conversation, whether the entity showed signs of emotions, using proper grammar). After every attribute, we also asked them about how that feature affected their opinion. ("To what degree did this feature influence opinion? 1 – *not at all*, 5 – *strongly affected*"). We also asked participants which type of error seems to be riskier and more serious for them: when a machine is mistaken for a human being (1) or when a human is mistaken for a machine (2).

And finally, they had to fill out *The Negative Attitudes Towards Robots Scale* (NARS, Nomura et al., 2006). The scale was developed to measure participants' anxieties toward robots. In previous studies the 14 item scale proved to be reliable (Cronbach α from .65 to .92; Nomura et al., 2006; Syrdal et al., 2009; Tsui et al., 2010). On our sample the reliability of the whole scale was high (Cronbach α = .85, subscale intercorrelations from .36** to .41**, so we used the average of all items as the indicator of the attitude of the participants.

Results

Most of the participants decided that the unknown entity was rather a robot (n = 484), 93 of them have chosen the "human" option, and 55 could not decide whether the entity was human or a robot (Table 3). We have found neither gender nor age-related differences.

Table 3

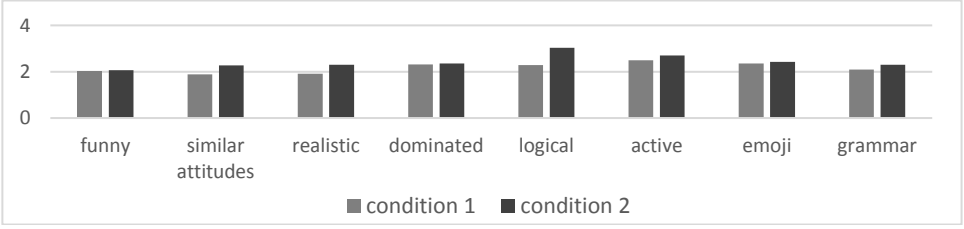
Descriptive Statistics of the Turing Type Decision in Study 2 (N = 632)

	1- I am sure it's a robot	2	3 - I cannot decide	4	5 - I am sure it's a human	Total
Frequency	195	289	55	81	12	632
Percentage	30.9	45.7	8.7	12.8	1.9	100

The averages of the Turing-type decision varied across the different conditions (Figure 2). The results showed that the highest difference between the two conditions was in the conditions of attitude difference, realistic conversation, logical conversation, and grammar.

Figure 2

The Averages of the Turing-Type Decision (1- I am Sure it is a Robot, 5 - I am Sure it is a Human) in the Different Conditions in Study 2



Note. Condition 1 covers the presence of the attribute (funny, similar attitudes, logical…) while Condition 2 means the attribute was missing (e.g. not funny, no similar attitudes, not logical…).

As we also wanted to test the predictive power of all the variables on the Turing-type decision, we conducted hierarchical multiple regression analyses (Table 5), and prior to it, the relevant assumptions of this statistical analysis were tested. The correlations between the predictors were not unacceptably high (Table 4). Predictors were divided into two distinct sets, to distinguish between the influence of demographics and the other variables. Block 1 contained age, gender, and NARS scores. Block 2 consisted of the 8 attributes (being funny, showing similar attitudes to the judge, how realistic were their answers, whether they dominated the conversation, how logical were their answers, whether they played an active role during the conversation, whether the entity showed signs of emotions, using proper grammar).

Table 4

Summary of Intercorrelations for the 8 Attributes in Study 2 (N = 632)

	1	2	3	4	5	6	7	8
1. funny	-							
2. attitudes	.12	-						
3. realistic	.16	.13	-					
4. dominated	.04	-.15	.10	-				
5. logical	-.07	.18*	.12	-.25**	-			
6. active	.11	.18	.20	.31	.14	-		
7. emotions	.18	.22*	-.14	.12	.21*	.16	-	
8. grammar	.06	.18*	.00	.02	.06	-.14	.19**	-
M	2.35	2.08	2.11	2.34	2.65	2.60	2.38	2.20
SD	1.21	1.17	1.34	1.37	1.32	1.35	1.19	1.22

Note. Attributes are: being funny, showing similar attitudes to the judge, how realistic were their answers, whether they dominated the conversation, how logical were their answers, whether they played an active role during the conversation, whether the entity showed signs of emotions, using proper grammar; $^{*}p < .05$; $^{**}p < .01$.

The regression revealed that at Step 1 just age contributed significantly to the regression model ($F(3, 628) = 5.679$, $p < .001$) but explained only 2.2% of the variance. Introducing the 8 attributes explained an additional 18.6% to the explanation of variance (R^2 change = .186; $F(3, 628) = 18.268$, $p < .001$), thus altogether 20.8% of the variance was explained by the model. Further results indicate that the Turing type of decision was most influenced by the variable realistic conversation ($\beta = .314$, $p < .01$), giving logical answers ($\beta = .115$, $p < .01$), and attitude similarity ($\beta = .097$, $p < .05$).

Table 5

Summary of Hierarchical Regression Analysis for Variables Predicting Turing-Type Decision (N = 632)

	Model 1			Model 2		
	B	*SE B*	β	*B*	*SE B*	β
gender	0.131	0.088	.058	0.077	0.081	.034
age	0.016	0.004	.149**	0.010	0.004	.096**
NARS	-0.065	0.076	-.033	-0.074	0.069	-.038
funny				0.061	0.060	.051
attitude				0.085	0.083	.097*
similarity						
realistic				0.242	0.227	.314**
dominant				0.043	0.029	.057
logical				0.089	0.031	.115*
active				0.001	0.029	.001
grammar				0.063	0.078	.081
emotions				0.214	0.202	.248
ΔR^2		.022**			.186**	
F for change in R^2		5.679**			18.268**	

*$p < .05$; **$p < .01$.

We wanted to compare the results from the regression to the self-reported importance of the attributes (Table 6). They were rather similar with some exceptions. Realistic conversations, logical answers, and attitude difference proved to be influential in both types of questions, while grammar counted more when asked explicitly and attitude similarity was more important according to the regression.

Table 6

Self-Reported Importance of the Attributes in Study 2 (N = 632)

	N	*M*	*SD*
logical answers	632	3.85	1.32
realistic conversation	632	3.73	1.34
proper grammar	632	3.58	1.23
active	632	3.53	1.06
dominant	632	3.41	1.43
similarity of attitudes	632	2.57	1.24
emotions	632	2.23	1.22
funny	632	2.08	1.21

To the question of which type of error seems to be riskier and more serious: 265 participants (42%) said that when a human is mistaken for a machine, while 367

(58%) have chosen the other option when a machine is mistaken for a human being. We have found neither gender nor age-related differences, the answer did not even depend upon the decision whether the entity was a human or a robot. But participants' negative attitudes towards robots proved to be a significant predictor in a logistic regression model ($\chi^2(1) = 10.349$, $p < .001$), meaning that those with a rather negative attitude had an odds of choosing a machine mistaken for a human as the more serious mistake of 1.621 times than the other Turing decision error.

Discussion of Study 2

We aimed to reveal the most influential factors in the decisions of the judges during the Turing test. Our participants functioned as "second judges", they had to read conversations between a judge and an entity and had to decide whether the entity was a human or a robot. The most influential variables were whether the conversation looked realistic enough, whether the entity gave logical answers, and whether it had similar attitudes to the perceiver. These findings are mainly in line with previous literature (Kleijn et al., 2019; Warwick & Shah, 2014, 2015). Though we have found some differences when asked explicitly compared to when analysing their answers by regression: while realistic looking conversations, logical answers, and attitude difference proved to be influential in both types of questions, grammar counted more when asked explicitly and attitude similarity was more important according to the regression.

Based on the findings of our first study we hypothesized that confusing a robot with a human being would appear as a riskier choice for the participants, which proved to be the case in our second study. Further research should clarify what are the underlying psychological mechanisms that can explain this phenomenon. Maybe one of the reasons is linked to our finding, namely that the more negative attitude we have toward robots, the worse we would feel to mistake a robot for a human being.

General Discussion

Artificial intelligence is one of the fastest developing areas of human technology, so to decide whether we are talking to a human or a robot is not just a theoretical question anymore, as chatbots are appearing in many areas of our social life.

Our two studies created a model of a situation that we already are familiar with: exchanging messages virtually with an entity, and the only information we have about this entity is the one it gives us through the messages. In general, the main contribution of our studies was to examine the importance of some social psychological factors that can play a role when judging in the Turing test. On the one hand, our results can contribute to our understanding of what constitutes humanness

and how aware are we of the motivators of our decision whether the other is a human or a robot. It turned out, that we expect humans to be imperfect (Kleijn et al., 2019), but we also expect an understanding of the rules of human communication. Violating these can result in dehumanization (mistaking a human for a robot). The maxim of relevance (we have to say relevant things during a conversation) (Grice, 1975) is one of the most sensitive ones stated Saygin and Cicekli (2002), after analysing Turing conversations. However, violating the maxim of manner (we have to be clear, brief, and have to avoid ambiguity) was rather interpreted as the unknown entity expressing its emotions, therefore a human trait. The violation of the maxim of quantity (we have to be as informative as we can, but no more than needed) by sharing too much information resulted in the impression of the unknown entity being a robot, but providing less information than required was not associated significantly to any of the decisions, and neither the validity of the information (maxim of quality). The results also added something to our social-psychological knowledge of dehumanization, particularly to the so-called mechanistic dehumanization, when humans are being likened to machines and denied characteristics of human nature, such as warmth and curiosity. The attributed characteristics usually are inertness, coldness, rigidity, fungibility, and lack of agency (Haslam, 2006).

On the other hand, this type of research is very important to understand the human-computer relationship, to design robots and chatbots that we are willing to collaborate with. One yet unanswered question could be whether we should design robots that resemble humans or it would rather cause a backlash.

We have to take the limitations of both studies into account when interpreting the results. Our research designs allowed us to conduct the examinations online, but at the same time, it became impossible to control many factors of the situation, such as the attention the participants paid. The majority of participants were Hungarians as we did not have access to enough native speakers, but the language of our study was English, and this could have also caused some difficulties, especially in judging the others based on their English grammar for example (even though we controlled the participants by making the advanced level of English a criterion for participation).

As human-computer interactions will play an important role in humanity's future, this type of research can give us an insight into the (maybe not-so-far) future, when we will be surrounded by machines and computer programs that have passed the Turing test.

Acknowledgments

Special thanks to Zsombor Hollay-Horváth, for developing the Turing research software.

Supported by the ÚNKP-18-2 New National Excellence Program of the Ministry of Human Capacities.

References

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin, 140*(6), 1556. https://doi.org/10.1037/a0037737

Block, N. (1981). Psychologism and behaviourism. *Philosophical Review, 40,* 5–43. https://doi.org/10.2307/2184371

Candello, H., Pinhanez, C., & Figueiredo, F. (2017). Typefaces and the perception of humanness in natural language chatbots. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3476–3487). https://doi.org/10.1145/3025453.3025919

Collins, H. M. (1990). *Artificial experts: Social knowledge and intelligent machines*. MIT Press.

Efferson, C., Lalive, R., & Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism. *Science*, *321*(5897), 1844–1849. https://doi.org/10.1126/science.1155805

French, R. M. (1990). Subcognition and the limits of the Turing test. *Mind*, *99*(393), 53–65. https://doi.org/10.1093/mind/XCIX.393.53

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (pp. 41–58). Brill. https://doi.org/10.1163/9789004368811_003

Halpern, M. (2006). The trouble with the Turing test. *The New Atlantis, 11,* 42–63.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, *1*(1), 43–54. https://doi.org/10.1007/BF00360578

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Hayes, P. J., & Ford, K. M. (1995). *Turing test considered harmful*. IJCAI.

Kleijn, De R., Wijnen, M., & Poletiek, F. (2019). The effect of context-dependent information and sentence constructions on perceived humanness of an agent in a Turing test. *Knowledge-Based Systems, 163,* 794–799. https://doi.org/10.1016/j.knosys.2018.10.006

Levin, L., & Peled, E. (2011). The attitudes toward prostitutes and prostitution scale: A new tool for measuring public attitudes toward prostitutes and prostitution. *Research on Social Work Practice*, *21*(5), 582–593. https://doi.org/10.1177/1049731511406451

Lortie, C. L., & Guitton, M. J. (2011). Judgment of the humanness of an interlocutor is in the eye of the beholder. *PloS One, 6*(9), e25085. https://doi.org/10.1371/journal.pone.0025085

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies, 7*(3), 437–454. https://doi.org/10.1075/is.7.3.14nom

Novak, D. W., & Lerner, M. J. (1968). Rejection as a consequence of perceived similarity. *Journal of Personality and Social Psychology*, *9*(2), 147–152. https://doi.org/10.1037/h0025850.

Saygin, A. P., & Cicekli, I. (2002). Pragmatics in human-computer conversations. *Journal of Pragmatics, 34*(3), 227–258. https://doi.org/10.1016/S0378-2166(02)80001-7

Saygin, A. P., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and Machines*, *10*(4), 463–518. https://doi.org/10.1023/A:1011288000451

Schwartz, S. H., & Struch, N. (1989). Values, stereotypes, and intergroup antagonism. In D. Bar-Tal, C. F. Graumann, A. W. Kruglanski, & W. Stroebe (Eds.), *Stereotyping and prejudice. Springer series in Social psychology*. Springer. https://doi.org/10.1007/978-1-4612-3582-8_7

Searle, J. R. (1980) Minds, brains, and programs. *Behavioural and Brain Sciences, 3,* 417–424. https://doi.org/10.1017/S0140525X00005756

Shah, H., & Henry, O. (2005). Confederate effect in human-machine textual interaction. *Proceedings of 5th WSEAS international conference on information science, communications and applications (*WSEAS ISCA) (pp. 109–114). Cancun, Mexico.

Shah, H., & Warwick, K. (2010). Testing Turing's five minutes, parallel-paired imitation game. *Kybernetes, 39*(3), 449–465. https://doi.org/10.1108/03684921011036178

Syrdal, D. S., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2009). The Negative Attitudes Towards Robots Scale and reactions to robot behaviour in a live Human-Robot Interaction study. *Journal of Biosocial Science, 1*(S1), 173–191.

Tsui, K. M., Desai, M., Yanco, H. A., Cramer, H., & Kemper, N. (2010). Using the "Negative Attitude toward Robots Scale" with telepresence robots. *Proceedings of the 10th performance metrics for intelligent systems workshop* (pp. 243–250).

Turing, A. M. (1950). Mind. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Warwick, K., & Shah, H. (2014). Good machine performance in Turing's imitation game. *IEEE Transactions on Computational Intelligence and AI in Games, 6*(3), 289–299. https://doi.org/10.1109/TCIAIG.2013.2283538

Warwick, K., & Shah, H. (2015). Human misidentification in Turing tests. *Journal of Experimental & Theoretical Artificial Intelligence, 27*(2), 123–135. https://doi.org/10.1080/0952813X.2014.921734

Weizenbaum, J. (1976) *Computer power and human reason.* W. H. Freeman.

Received: May 17, 2021

Appendix

Structure of the Turing Research Software

The connections between the parts of the application were the following: the TCG has been run ahead of the test sessions, the generated codes were presented to the SSA. The CSA was only communicating with the SSA layer through a secure connection (HTTPS for static content, WSS or Secured WebSocket for real-time communication). The SSA managed the communication between the SSA and the DBA layers. The CSA and the DBA did not have any direct connections. As for the implementation and business logic, the TCG was implemented in Python. It generated 5 letters long random UIDs from the set of lower and uppercase letters from the English alphabet and any digits, but the characters 0 (zero), o (lowercase O), O (capital O), 1 (digit one), I (capital i) and l (lowercase L) were excluded from the set to minimize the chance of misreading a UID. The CSA layer was implemented as a website. After the page was loaded, it opened the real-time secure connection to the SSA and managed the flow of the entire research process. If the real time connection encountered any kind of error, it automatically fixed the connection. If fixing the connection was not possible (e.g. the participant went offline) it displayed a loading bar until the connection was established again. The SSA layer was an event-driven server implemented in NodeJS. As the CSA sent data to the SSA, the SSA decided what to do with that information and made the corresponding action (validating UID, sending messages to the other participant on the chat, saving questionnaire results, etc.) and stored the results on the DBA. The DBA was a MongoDB engine that was only used for storing the results, it was not containing any logic on its own.