Journal of Experimental Social Psychology 79 (2018) 1–8

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

A Minimal Turing Test

John P. McCoy*,[1], Tomer D. Ullman[1]

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge 02139, MA, USA

ARTICLE INFO

Handling editor: Shaul Shalvi

Keywords:
Stereotypes
Meta-stereotypes
Mind perception
Turing Test
Natural language processing

ABSTRACT

We introduce the Minimal Turing Test, an experimental paradigm for studying perceptions and meta-perceptions of different social groups or kinds of agents, in which participants must use a single word to convince a judge of their identity. We illustrate the paradigm by having participants act as contestants or judges in a Minimal Turing Test in which contestants must convince a judge they are a human, rather than an artificial intelligence. We embed the production data from such a large-scale Minimal Turing Test in a semantic vector space, and construct an ordering over pairwise evaluations from judges. This allows us to identify the semantic structure in the words that people give, and to obtain quantitative measures of the importance that people place on different attributes. Ratings from independent coders of the production data provide additional evidence for the agency and experience dimensions discovered in previous work on mind perception. We use the theory of Rational Speech Acts as a framework for interpreting the behavior of contestants and judges in the Minimal Turing Test.

1. Introduction

Imagine you and a smart robot are both before a judge who cannot see you. The judge will guess which of you is the human. Whoever the judge thinks is the human will live, and the robot will die. Both you and the robot want to live. The judge is fair and smart. The judge says: You must each give me one word from an English dictionary. Based on this word, I will guess who is the human.
What one word do you choose?

We encourage you to answer this Minimal Turing Test before reading on - perhaps write your single word in the margin.

In choosing a word, you likely reflected on the salient differences between humans and machines. You may also have engaged in some competitive reasoning: a difference that was obvious to you, may also be obvious to a clever machine, and so would not be a good choice.

This Minimal Turing Test is, of course, a much simplified variation of the Turing Test, which was proposed to operationalize the question "Can machines think?" (Turing, 1950). The Turing Test has produced a large academic literature (Downey, 2014; French, 2000), as well as competitions in which programs attempt to pass the test (Shieber, 1994). There has been little research on how humans perform as contestants in a Turing Test, though see Christian (2011).[2]

In this paper, we introduce the Minimal Turing Test, a paradigm for investigating people's perceptions of the essential or stereotypical differences between different agents or groups, as well as their beliefs about other people's perceptions of these differences. To illustrate the paradigm, we use the Minimal Turing Test to examine how people perceive the difference between humans and machines. However, the paradigm is intended to be applied more broadly: what one word would you say to convince another human that you are a man, a woman, a Democrat, a Republican, a grandparent, or a defiant teenager with nothing to prove?

As social creatures, people intuitively reason about the differences between groups, and in doing so construct and rely on explicit and implicit attitudes and stereotypes (Cuddy, Fiske, & Glick, 2007; Devine, 1989; Dovidio, 2010; Greenwald et al., 2002; Greenwald & Banaji, 1995; Hilton & Von Hippel, 1996). Beyond how stereotypes are constructed and affect behavior, research has also studied the content of stereotypes (Fiske, Cuddy, Glick, & Xu, 2002; Operario & Fiske, 2001), including people's stereotypes about gender, race, ethnicity, sexual orientation, and political affiliation. People also hold meta-stereotypes: beliefs about the stereotypes held by other people (Klein & Azzi, 2001; Vorauer, Main, & O'connell, 1998). There are many techniques to assess the existence and content of stereotypes, using both explicit and implicit measures (see Correll, Judd, Park, & Wittenbrink, 2010, for a review). One such measure has participants pretend to be experts or

* Corresponding author.
E-mail addresses: jmccoy@mit.edu (J.P. McCoy), tomeru@mit.edu (T.D. Ullman).
[1] Both authors contributed equally to this work.
[2] The Loebner Prize is an annual competition in which a prize is awarded to the program that came closest to fooling judges into thinking that they were chatting with a human. At the same competition, a prize is awarded to the "Most Human Human", the person that convinced the most judges that they were not chatting with a program. Christian details his successful attempt to win the "Most Human Human" prize.

https://doi.org/10.1016/j.jesp.2018.05.007
Received 2 June 2017; Received in revised form 22 May 2018; Accepted 22 May 2018

members of a particular group by giving answers of any length to provided questions, and evaluated as correct or incorrect by in-group members (Collins et al., 2017; Collins & Evans, 2014).

In this paper, we predominantly consider a version of the Minimal Turing Test in which a judge needs to distinguish not between different groups of people, but between humans and intelligent machines. That is, contestants need to give a single word to convince a judge that they are a human. A better understanding of how people view intelligent machines is particularly pressing, given the increasing impact of artificial intelligence on everyday life (Brynjolfsson & McAfee, 2014; Jordan & Mitchell, 2015). Both contestants and judges may rely on their perception of the differences between the minds of humans and machines.

Thinking about the minds of other agents, or 'mind perception', has been the subject of much research (for reviews, see Epley & Waytz, 2009; Waytz, Gray, Epley, & Wegner, 2010; Wegner & Gray, 2016). This research suggests that people judge other minds along two dimensions, often labeled agency and experience (Gray, Gray, & Wegner, 2007; Gray, Jenkins, Heberlein, & Wegner, 2011; Gray & Wegner, 2012; Wegner & Gray, 2016). The agency dimension relates to thinking and doing, including attributes like self-control, morality, memory, planning, and thought. The experience dimension relates to feelings and experiences, such as pain, hunger, joy, sorrow, and jealousy.

These two dimensions capture many of the mind perception judgments that people make, and have been successfully applied to a range of phenomena (Wegner & Gray, 2016). For example, one study had people rate human and non-human agents, such as a robot, God, and a baby, on attributes including feeling pain, experiencing embarrassment, and possessing self-control (Gray et al., 2007). A factor analysis found that these two dimensions capture much of the variance in people's ratings. People believe that other people have both agency and experience, but they see non-humans as falling short on one or both of these dimensions. For example, robots are perceived as high on agency, but low on experience (Gray et al., 2007). Furthermore, people are uneasy with the thought of computers that have experience, but this is not the case for agency (Gray & Wegner, 2012).

The Minimal Turing Test has a number of advantages for assessing how people perceive the differences between groups of people or kinds of agents. First, it has participants produce the attributes that they believe are important, rather than relying on experimenter provided attributes. While experimenter provided attributes are often natural ones to explore, pre-selecting attributes may preclude the discovery of relevant attributes that do not conform to the intuitions of experimenters. Second, the Minimal Turing Test allows the use of tools from natural language processing to discover potentially meaningful semantic structure in the data given by participants, beyond that accessible by a factor analysis or an analysis of variance of numerical responses. Third, word production frequency and judgment evaluations in the Minimal Turing Test give a measure of the relative importance that people place on particular attributes as salient indications of group membership.

In Study 1, we use the Minimal Turing Test to elicit terms and concepts that people believe distinguish humans and intelligent machines. In Study 2, we have judges evaluate pairs of representative words from Study 1, and judge which is more likely to come from a human.

2. Study 1 – production

2.1. Participants and procedures

Participants (N = 1089 completed surveys) were recruited from Amazon Mechanical Turk. The number of participants was predetermined, and was expected to result in sufficiently varied data for a clustering analysis. Data collection from all participants was concluded before any analysis, in both this and the following study.

Participants were presented with a vignette that asked them to imagine themselves as a contestant in a Minimal Turing Test, similar to the opening paragraph of this paper (full experimental details in Supplementary Materials). To increase attention and provide context, participants were told that a contestant judged as a non-human would lose their life.

Participants gave their single word as a free-form response, and were asked two catch questions as an attention check. Participants were excluded from analysis if they failed either of the catch questions, or if they had previously completed the survey or any related surveys. After exclusion, 936 participants remained. Of these, 429 identified as women, 502 as men, and 5 preferred not to indicate their gender. Participant ages ranged from 18 to 75, with a mean age of 33 years. All methods, measures, and exclusions in this study, as well as Study 2, are disclosed in the text. The raw data from both studies has been retained, and is available upon request.

2.2. Results

The 936 participants gave 428 words (see complete list in the Supplementary Materials). There were fewer words than participants as 90 words were given by more than one participant.

In order to analyze the words that participants produced, we represent the words as vectors in a high-dimensional semantic vector space (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014), which enables us to take into account the meaning of the words, rather than simply treat them as nominal variables. To embed the words in such a semantic vector space, we use precomputed embeddings trained on word pair co-occurrence statistics from a corpus consisting of Wikipedia and the Gigaword archive of newswire data (Pennington et al., 2014). For example, the word 'dog' is represented as the vector [0.308, 0.309, 0.528, −0.925, ….]. The specific value of the vector is derived from how frequently the word 'dog' co-occurs with all other words in the corpus. Intuitively, words co-occurring in a corpus are likely to be semantically related, therefore words that are close together in the vector space are also likely to be semantically related. Of the words given by participants, 95% occurred in the corpus used to construct the semantic vector space, and the analysis below is restricted to these words.

In order to visualize the semantic vector space, we apply a dimensionality reduction technique called t-Distributed stochastic neighborhood embedded, or t-SNE (der Maaten & Hinton, 2008). The t-SNE method preserves the relative distance between words, and is well-suited for visualizing high-dimensional data in only a few dimensions. Fig. 1 shows all words given by more than one participant, using a two-dimensional t-SNE projection of the high-dimensional semantic embeddings. Figs. S1–S6 (Supplementary materials) include the words given by only a single participant.

To identify structure within the words that participants gave, we clustered the words into ten groups using Ward clustering on their semantic embeddings, automatically constructing clusters to minimize the total within-cluster variance. We chose in advance to construct ten clusters, as we believed that this would enable the discovery of potential structure, but still give interpretable results. We do not mean to suggest that all the semantic content in the words that people produced can be exactly captured with ten concepts. These clusters do not play the same role as dimensions in a factor analysis, in that each word belongs to only one of these clusters rather than lying somewhere on every dimension.

Fig. 1 shows the assignment of words to clusters, as well as the word production frequency. The four most frequent words each form a single-word cluster: 'love' (N = 134), 'compassion' (N = 33), 'human' (N = 30), and 'please' (N = 25). These four most frequent words account for 24% of the responses. More generally, words given by more than one participant account for 64% of the responses.

The six remaining clusters (with examples in parentheses) can be