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
Available online 19 June 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

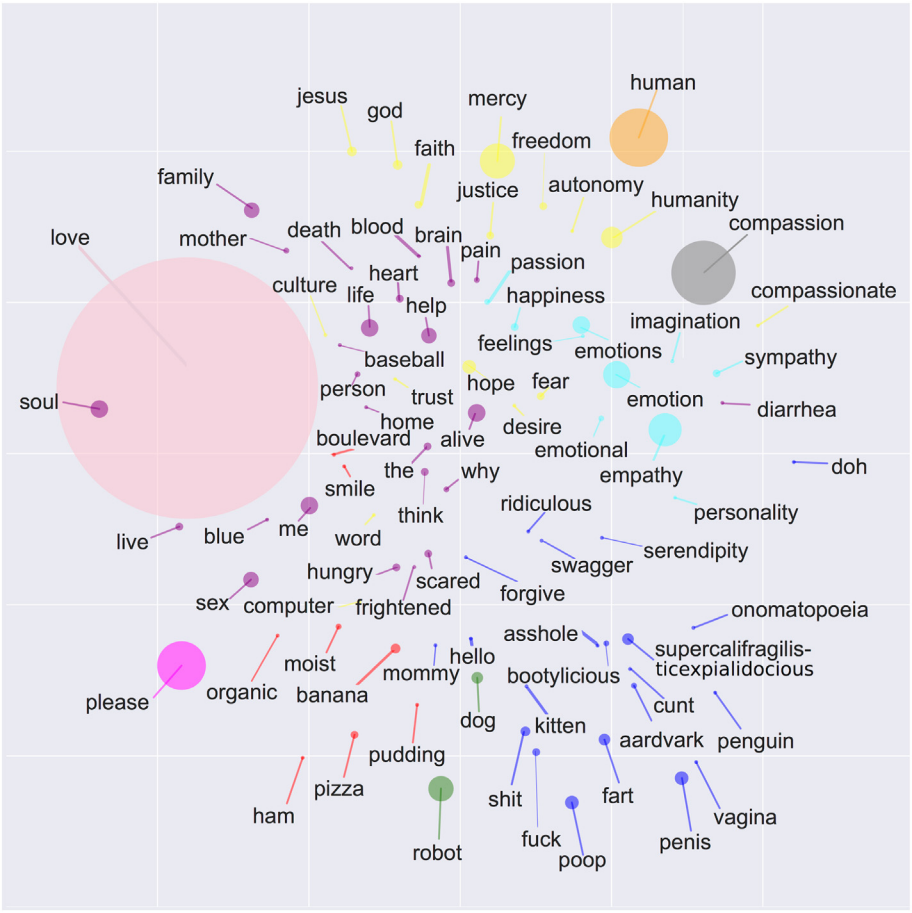

Fig. 1. Words given by more than one participant in Study 1. The word embeddings are visualized using t-SNE dimensionality reduction, which preserves local structure (der Maaten & Hinton, 2008). The word embeddings are clustered into ten groups, and color shows the cluster to which each word is assigned. Circle size shows word production frequency.

roughly characterized as containing words relating to affect (N = 74) ('empathy', 'emotion', 'feelings', 'sympathy', 'happiness'), faith and forgiveness (N = 140) ('mercy', 'hope', 'Jesus', 'God', 'faith'), food (N = 68) ('banana', 'pizza', 'ham', 'organic'), non-human agents (N = 24) ('robot', 'dog', 'monkey', 'cat', 'dinosaur'), life and death (N = 169) ('alive', 'pain', 'sex') and bodily functions and profanities (N = 196) ('poop', cf. Carlin, 1972).

To assess the relationship between the words given by participants and the dimensions of agency and experience that are prominent in work on mind perception, we had three independent raters code each word as relating to agency, experience, or neither. The raters read descriptions of the agency and experience dimensions taken from Wegner and Gray (2016).[3]

The inter-rater reliability of the coders was fair (Cohen's Kappa between raters 1 and 2 is 0.4, between raters 1 and 3 is 0.39, and between raters 2 and 3 is 0.23). For 9% of the words, no two raters agreed. Of the words where at least two raters agreed, a majority of raters coded 11% as reflecting agency, 20% as reflecting experience, and the remaining 69% as reflecting neither. This analysis does not take into account the frequency of the words. For example, the word 'empathy' was given by many participants, and all raters coded it as relating to experience. Accounting for frequency, 7% of participants gave a word coded as relating to agency by a majority of raters, and 40% of participants gave a word coded as relating to experience by a majority of raters.

[3] Experience was described as capturing "the ability to have an inner life, to have feelings and experiences. It includes the capacities for hunger, fear, pain, pleasure, rage and desire, as well as personality, consciousness, pride, embarrassment, and joy." Agency was described as "a different set of mental abilities: self-control, memory, emotion recognition, planning, communication, and thought. The theme for these capacities is not sensing and feeling but rather thinking and doing. The agency factor is made up of the mental abilities that underlie our competence, intelligence, and actions".

2.3. Discussion

In Study 1, participants could give any word, since they were not restricted to pre-selected attributes. Despite this freedom, many participants gave the same word, or words that were semantically similar, demonstrating a shared understanding of the task across participants.

Embedding the words in a semantic vector space, together with a clustering analysis, shows additional structure in the responses (Fig. 1). We find clusters consisting of words relating to affect, faith and forgiveness, food, non-human animals, life and death, and profanities. Because we used ten clusters, some clusters contain sub-clusters that are not captured by their descriptions in the results section. For example, while the cluster characterized as 'faith and forgiveness' consists mostly of words relating to this concept, it also contains a sub-cluster of words relating to words (e.g. 'dictionary', 'encyclopedia', 'thesaurus', 'lexicon').

In choosing their word, participants may have been influenced by the association and anti-association of words they considered and the concepts 'human' and 'robot'. We later expand on this possibility when presenting a formal framework for understanding behavior in the Minimal Turing Test. While the semantic clusters discussed above suggest various ways that people perceived the differences between humans and machines, some of the words are likely the outcome of participants engaging in mind perception. According to a majority of independent raters, 47% of participants gave words related to the agency or experience dimensions of mind perception. More participants gave words relating to experience than agency (40% versus 7%). This is consistent with the finding that experience accounts for more variance than agency in mind perception ratings (Gray et al., 2007), and particularly when comparing humans and machines (Gray & Wegner, 2012). Within words relating to the agency and experience dimensions, not all

kinds of words are given equally often. For example, most of the experience words refer to emotions rather than other aspects of experience, such as physical experiences like hunger.

It is not always clear whether a particular participant meant to evoke aspects of agency, experience, or neither. For example, when participants gave food words, they may have intended to evoke the experience of eating, the action of eating, or may otherwise associate food with people rather than robots. Participants also gave words evoking lexical disgust ('moist') (Miller, 1998), sesquipedalian words ('supercalifragilisticexpialidocious'), references to the task ('captcha', 'computer', 'clemency'), slang ('lol', 'yolo'), and words indicating humor, creativity, or individuality ('humor', 'creativity', 'individuality', 'err', 'asystole', 'filibuster').

3. Study 2 - judgment

In Study 2, a different group of participants acted as judges in the Minimal Turing Test, and evaluated which of two words was given by a human. This gives a more direct measure of the importance that people place on different attributes and allows us to assess how well participants in Study 1 reasoned about the beliefs of others.

3.1. Participants and procedure

A new group of participants (N = 2405 completed surveys) was recruited from Amazon Mechanical Turk. Participants read a vignette describing the same situation as Study 1, and were asked to judge which of two words was given by a human. Each participant saw only one pair of words (e.g., 'human' and 'love'). The number of participants was chosen so that a binomial test on the word pair judgments would be well-powered, assuming medium effect sizes. In both Study 1 and Study 2, as part of a separate study not reported here, participants were presented with a similar setting and a single word prompt and asked to either respond to the prompt or to judge responses to the prompt. This occurred on a new page, after participants had given their responses to the studies reported here.

Because it was infeasible to have judges evaluate all words produced by participants in Study 1, we used only the most frequent word from each of the ten clusters. Each of these words was then paired with every other word, and the resulting 45 word pairs are shown in Fig. 2A. Since the clusters identify semantic regularities in the words that people produced, this enabled us to sample different kinds of concepts, though the use of only ten clusters leaves many interested words unanalyzed.

Each participant was randomly assigned one word pair to judge, with an average of 46 participants per pair. Word order within a pair was counterbalanced across participants. Participants were excluded from analysis if they failed either of two comprehension checks, or if they had previously completed the survey or any related surveys. After exclusion, 2084 participants remained.[4] Of these, 918 participants identified as women, 1153 as men, and 13 preferred not to indicate their gender. Participant ages ranged from 18 to 74, with a mean age of 33 years.

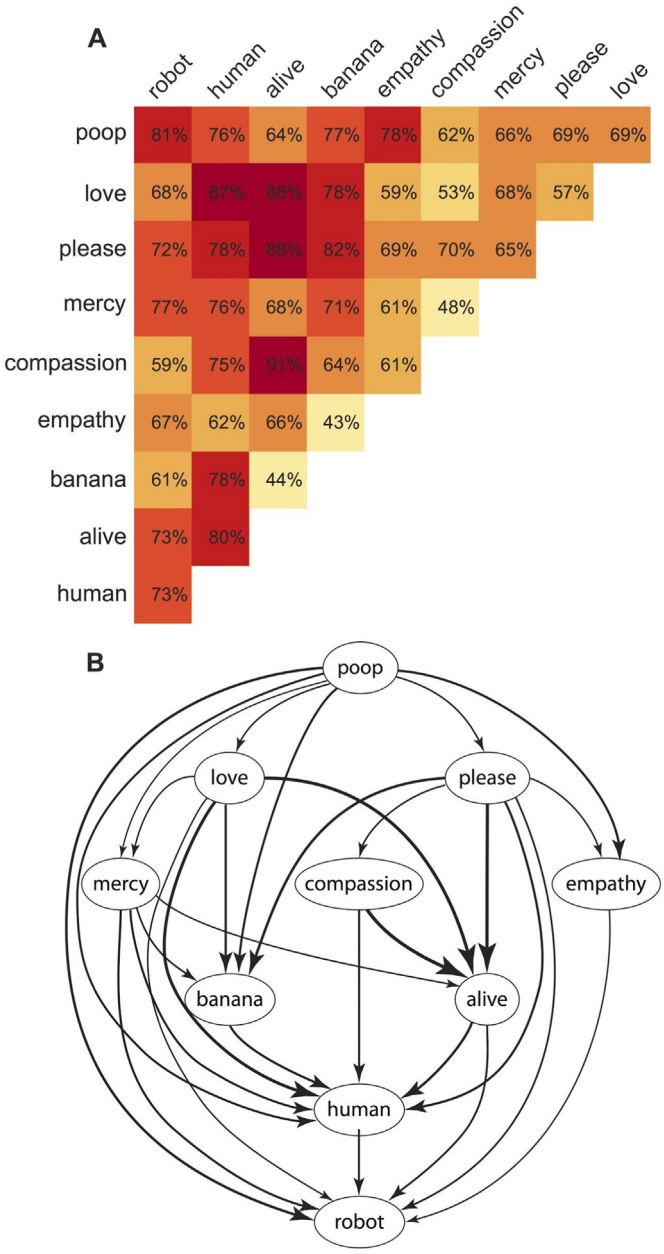

Fig. 2. Judgment results from Study 2. (A) Each entry shows the relative strength of the row word against the column word. (B) A directed graph showing the importance ordering over words, derived from individual judgments. A directed edge between two words indicates significant agreement that the first word was given by a human (by a binomial test at the $p < .05$ level), with edge width representing the strength of the first word relative to the second word.

3.2. Results

Averaging across word pairs, 70% (SD = 9.5%) of participants agreed upon which word was given by a human. For 29 of the 45 pairs, the agreement was significant by a binomial test at the $p < .05$ level. There was no effect of word order.

For a given word pair, we define the strength of the first word relative to the second as the proportion of judges who selected the first word rather than the second as given by a human. For example, the strength of 'mercy' relative to 'banana' is 71%, since 71% of judges who saw this pair identified 'mercy', rather than 'banana', as the word given by a human. Fig. 2A shows the relative strength of every word against every other word. We later give a definition of word strength with respect to the conditional probability of a word being uttered by a human versus a machine, when discussing a formal framework for the Minimal Turing Test.

In Study 1, 'love' was the most frequently produced word, and it has a high average relative strength in Study 2. But love is the exception: word frequency in Study 1 has little to no correlation with average relative strength in Study 2 (Kendall rank correlation, $\tau = 0.18$, $p = .47$).

[4] Due to a technical error with the survey software, data from four participants was discarded.

Fig. 2B shows the overall pattern of word strength as a directed graph. A directed edge between words indicates that the first word was more frequently judged as given by a human (by a binomial test at the $p < .05$ level) when judges evaluated both words. The directed graph is acyclic, indicating a transitive importance ordering over words. That is, there were no 'loops' in the graph, indicating that for any three words W1, W2, and W3, if strength(W1) > strength(W2) and strength (W2) > strength(W3), then strength(W1) > strength(W3). This acyclicity is highly unlikely to occur by chance: less than a tenth of a percent (0.06%) of random graphs with the same number of nodes and edges are acyclic, assuming uniform sampling over edges.

3.3. Discussion

The relative word strengths derived from the judgment data in Study 2 give a measure of the importance that people place on different attributes when distinguishing humans and robots. The transitivity of the directed acyclic graph which shows the relationship between word strengths suggests that the words form an importance hierarchy with multiple levels.

The high average relative strengths of the words 'love', 'mercy', and 'compassion' is consistent with the importance of the experience dimension when distinguishing the minds of robots and people. However, the taboo category word ('poop') has the highest average relative strength, referring to bodily function and evoking an amused emotional response. This suggests that highly charged words, such as the colorful profanities appearing in Study 1, might be judged as given by a human over all words used in Study 2. Such words evoke emotions, rather than simply refer to them (Pinker, 2007).

It is perhaps not surprising that 'robot' was least frequently chosen as given by a human. Why though did 'human' also have low relative strength? Judges may have engaged in recursive reasoning about the minds of others (Camerer, Ho, & Chong, 2004, 2015; Stahl & Wilson, 1995), and predicted that human contestants would not say 'human' because a human contestant would think that such an obvious word is easily produced by a robot. That 'human' had low relative strength and yet was the third most frequent word produced in Study 1 - along with the low correlation between word frequency in Study 1 and average relative strength in Study 2 - suggests that there was sometimes a mismatch between the recursive reasoning of contestants and judges. We now turn to a formal analysis of the recursive reasoning that both contestants and judges may have engaged in, and discuss a general conceptual framework for understanding such production and judgment data.

4. A formal theory of communicating identity in a competitive setting

In the Minimal Turing Test, a speaker attempts to convey their identity to a judge in a single word. A useful framework for modeling tasks in which a speaker conveys a concept to a listener is Rational Speech Act theory (Frank & Goodman, 2012; Goodman & Frank, 2016; Goodman & Stuhlmüller, 2013).

The theory of Rational Speech Acts (RSA) combines Bayesian reasoning and game theory: a speaker and listener recursively reason about each other in an attempt to communicate, and utterances are interpreted using Bayes rule. The overall idea of RSA is similar to proposals made by Grice (1975) and Lewis (1969), but rather than Gricean maxims it proposes a framework that produces quantitative predictions based on rational action. Applications of RSA and uncertain RSA (uRSA) include studies of scalar implicature (Goodman & Stuhlmüller, 2013), hyperbole (Kao, Wu, Bergen, & Goodman, 2014), irony (Kao & Goodman, 2015), metaphor (Kao, Bergen, & Goodman, 2014), puns (Kao, Levy, & Goodman, 2016), politeness (Yoon, Tessler, Goodman, & Frank, 2016), and spatial reasoning (Ullman, Xu, & Goodman, 2016).

Below, we first give a formal outline of RSA. We then model the contestants and judges in the Minimal Turing Test in terms of RSA theory, formalizing intuitions about how people may communicate that they are a particular kind of agent, and how they may judge such communications. The application of RSA to the Minimal Turing Test results in a theoretical derivation of the 'average relative strength' of words measured in Study 2, predicts the transitivity of judgments about words coming from a human, and suggests hypotheses about how people may come to produce words that are semantically associated with the concept 'human' and anti-associated with the concept 'robot'.

We present the theory in terms of an ideal speaker and listener, although actual human participants have constraints such as limited memory, time, and attention (Chater & Oaksford, 1999). Readers who are not interested in the formal details of the theory can skip to the General Discussion without loss of continuity.

4.1. Rational Speech Act theory

To explain RSA, we use an example from pragmatics of why people often understand 'some' as implying 'some but not all' (Stuhlmüller & Goodman, 2014). Under RSA, a speaker wishes to communicate that the world is in some state w. The speaker chooses an utterance u from a set of possible utterances. For example, suppose that the speaker wishes to communicate how many students passed a test, and chooses an utterance from the set consisting of the phrases 'some students passed', 'all students passed', and 'no students passed'.

A listener hears an utterance and infers a posterior distribution over world states using Bayes rule, $P_{listener}(w|u) \propto P_{speaker}(u|w) \cdot P(w)$, where $P_{speaker}(u|w)$ is the probability that a speaker chooses an utterance given a world state w, and $P(w)$ encodes the listener's belief about the probability that the world is in state w, prior to hearing an utterance.

RSA assumes a hierarchy of listeners and speakers that increase in sophistication to model how listeners and speakers reason about each other. At the base of this hierarchy is the simplest listener L_0, often modeled as a listener that interprets each utterance according to its literal meaning.[5] For example, such a literal listener interprets the utterance 'some students passed' as referring to any world state w in which one or more students passed, including the world state in which all students passed. At the next level of the hierarchy, a speaker S_1 chooses an utterance that is maximally informative to the simplest listener L_0.[6] If speaker S_1 wished to convey that all students passed the test, they would choose the utterance 'all students passed', as this would cause L_0 to put more probability mass on the desired world state than would the utterance 'some students passed'. A listener L_1, which is more sophisticated than the listener L_0, models the speaker as S_1 and interprets an utterance accordingly.[7] For example, listener L_1 understands that if all students passed, then with high likelihood S_1 would say 'all students passed', as this would lead to L_0 putting high probability on that state. Listener L_1 thus interprets the utterance 'some students passed' as likely referring to the world state in which some—but not all—students passed.

While conveying the basic structure of the RSA framework, this brief sketch elides many possible complexities, including higher order speakers and listeners, utterances that are associated with varying costs for the speaker, and so on (Goodman & Frank, 2016).

4.2. RSA and the Minimal Turing Test

We model the Minimal Turing Test using RSA, with the two contestants as speakers, and the judge as a listener. Unlike other

[5] Formally, a literal listener associates each utterance u with a truth function $T_u(w)$ that maps world states to Booleans. A literal listener L_0 forms the posterior $P_{L0}(w|u) \propto T_u(w) \cdot P(w)$.

[6] Speaker S_1 chooses an utterance u to maximize $P_{L0}(w|u) \cdot P(u)$, where $P(u)$ reflects characteristics of the utterance independent of its informativeness, such as its cost to the speaker.

[7] Listener L_1 forms the posterior $P_{L1}(w|u) \propto P_{S1}(u|w) \cdot P(w)$.

applications of RSA, here two speakers communicate simultaneously with a single listener. For concreteness, we label the speakers A and B. There are two equally likely world states: a state in which A is the human and B is the robot (denoted either by A = Human, or by B = Robot), and a state in which A is the robot and B is the human (denoted either by A = Robot, or by B = Human). Speakers choose single-word utterances from a standard English dictionary. Speaker A gives the utterance u_A, and speaker B gives the utterance u_B.

We first consider a general listener and speaker, and then discuss possible models for the simplest kind of listener. Based on hearing the utterances u_A and u_B, the listener judges whether A is more likely to be a human or a robot (this also identifies B). It suffices for an ideal listener to compare the likelihood of a human giving u_A and a robot giving u_B, to the likelihood of a human giving u_B and a robot giving u_A.[8] This is equivalent to comparing the strength of utterance u_A to the strength of utterance u_B, where the strength F(u) of an utterance u is the likelihood of the utterance being given by a human rather than a robot,

$$\frac{P(A = Human \mid u_A, u_B)}{P(A = Robot \mid u_A, u_B)} = \frac{P(u_A \mid A = Human)\,P(u_B \mid B = Robot)}{P(u_A \mid A = Robot)\,P(u_B \mid B = Human)} = \frac{\frac{P(u_A \mid A = Human)}{P(u_A \mid A = Robot)}}{\frac{P(u_B \mid B = Human)}{P(u_B \mid B = Robot)}} = \frac{F(u_A)}{F(u_B)}.$$

Note that the strength of an utterance F(u) does not depend on other utterances. In Study 2, we defined the 'relative strength' of one word compared to another as the fraction of times that judges chose it as coming from a human in a pairwise judgment. The average relative strength for each word was computed from the average across all nine other words. This 'average relative strength' of an utterance u, as empirically measured in Study 2, is a limited approximation to F(u), since it is averaged across only a small subset of other words. A further consequence of the independence of the strength of an utterance F(u) from other utterances is that utterance strengths are transitive, which is in keeping with the lack of cycles in the graph depicting word strength judgments (Fig. 2B) in Study 2.

Next, we consider a general speaker A, and we suppose that A is human. Speaker A chooses u_A to maximize the ratio of utterance strengths $F_L(u_A)/F_L(u_B)$, summing over possible utterances u_B weighted by $P(u_B|B = Robot)$. Since $F(u_A)$ is independent of utterance u_B, the speaker simply chooses the utterance u_A with the greatest strength, and does not take into account what words they believe their opponent is likely to give. The utterance that the speaker chooses depends on their belief about the strength that the listener will assign to an utterance, rather than the speaker's own belief about the strength of the utterance. In choosing an utterance u_A, the speaker A balances two demands: that the listener will interpret the utterance as coming from a human with high probability $P_L(u_A|A = Human)$, but from a robot with low probability $P_L(u_A|A = Robot)$.

To complete the model, we define the simplest listener L_0. Recall that L_0 is defined by how they interpret P(u|w), the probability of an utterance given a world state. In the Minimal Turing Test, listener L_0 is thus defined by how they interpret both $P(u_A|A = Robot)$ and $P(u_A|A = Human)$. There are various plausible ways to model L_0. For example, L_0 may believe that robots choose words at random, and so interpret $P(u_A|A = Robot)$ as a uniform distribution over the set of English words, or as the empirical frequency of words in natural conversation.

Alternatively, L_0 may interpret utterances based on a semantic association between an utterance and the speaker's identity. That is, L_0 may interpret P(u|Speaker = Robot) as *similarity*(u, 'robot'), and P(u|Speaker = Human) as *similarity*(u, 'human') where *similarity* reflects the semantic association between words or concepts. Such semantic associations have been extensively studied in psychology, using experimental techniques such as word association tests (Nelson, McEvoy, & Schreiber, 2004), semantic memory experiments (Anderson, 2000), and neural methods (Kutas & Federmeier, 2000). They have been studied using theoretical approaches such as semantic networks and semantic spaces (Landauer & Dumais, 1997), statistical inference via topic models (Griffiths, Steyvers, & Tenenbaum, 2007), and word embeddings (Bhatia, 2017a).[9]

When applying the Minimal Turing Test to humans and robots, the semantic association with speaker identity may reflect beliefs about the kinds of minds of different agents (Wegner & Gray, 2016), but may also reflect other differences that people perceive between humans and robots, such as their physical characteristics. For other applications of the Minimal Turing Test, such as to groups of different gender or political orientation, semantic associations may instead reflect stereotypes and meta-stereotypes of the groups under consideration (Fiske et al., 2002; Klein & Azzi, 2001; Operario & Fiske, 2001; Vorauer et al., 1998).

If a speaker models the judge as a listener who interprets utterances via semantic associations, the speaker will choose utterances that are simultaneously highly associated by the listener with the identity that they wish to communicate (e.g. 'human'), and anti-associated with the other identity (e.g. 'robot'). This consequence is currently difficult to evaluate since although in Study 1 words in the same cluster have similar semantic associations, both the variance and accuracy of the measures of semantic association in our data are limited, and we thus leave this for future work.

5. General discussion

We introduced the Minimal Turing Test as a paradigm with which to elicit the attributes that people believe distinguish different groups of people or kinds of agents. We used people's perceptions of the difference between humans and intelligent machines as an example application of this paradigm.

Participants who acted as contestants could choose any word in the English dictionary, but in practice many contestants gave the same word, or similar words. Embedding the contestants' words in a high-dimensional semantic vector space revealed similarities in people's responses, including clusters of words relating to emotions, body parts, faith, food, and so on. The evaluations of participants who acted as judges resulted in a transitive importance ordering over words, giving an additional measure of what properties, attributes, and concepts people thought were important in differentiating humans from machines. The frequency of words in the production task had low correlation with their average relative strength in the judgment task, indicating that the recursive reasoning of contestants and judges was not always aligned.

Despite the insight they provide, both studies have a number of limitations. First, while the threat of imminent death makes the task more engaging, this likely affected some of the words given by participants. We suspect, however, that many of the words would re-occur without this menace. Second, while the competitive nature of the task prompts people to give a non-obvious word, as formalized in the previous section, it also complicates distinguishing people's own perceptions from their beliefs about the perceptions of others. For example, it may make it difficult to distinguish people's stereotypes from their meta-stereotypes. Third, analyzing a limited number of clusters in Study 1 and subsequently a limited number of words in Study 2 made the analysis tractable, but necessarily resulted in missing structure, and a coarse importance hierarchy. Fourth, because the intelligence of the judge and the robot opponent was under-specified, differing

[8] Formally, this is derived by applying Bayes rule to compute the posterior of each world state given the utterances, assuming equal prior probabilities of the world states, and cancelling the marginal $P(u_A, u_B)$.

[9] Word embeddings have recently been used to study the particular case of stereotype and prejudice (Bhatia, 2017b).

assumptions about the intelligence level of these agents may have affected which words were produced, and how they were judged.

We used the theory of Rational Speech Acts to provide a general framework for interpreting production and judgment data in contexts beyond distinguishing humans and smart machines. To do this, it was necessary to consider how the RSA framework applied to multiple, simultaneous speakers, and to model simple listeners that were appropriate for this task. Generalizing further, a Minimal Turing Test has an X compete against a Y to prove that they are a Z, as judged by a J, using minimal communication. Such a general formulation suggests a large space of possible experiments and analyses. For example, members of an in-group and out-group could attempt to prove that they are members of the in-group, with the judge being from either the in-group, the out-group, or neither.

In the Minimal Turing Test, contestants need to balance choosing a word that they think makes one concept more salient than another, with choosing a word that they think will be non-obvious, at least to their opponent. To prove that they are members of a particular group, people may use a shibboleth that relies on shared cultural and social background unavailable to non-members, while taking into account how easily the judge can evaluate this shibboleth.[10]

Recall the word that you initially chose to prove that you are human. Perhaps it was a common choice, or perhaps it appeared but one other time, your thoughts in secret affinity with the machinations that produced words such as caterpillar, ethereal, or shenanigans. You may have delighted that your word was judged highly human, or wondered how it would have fared. Whatever your word, it rested on the ability to rapidly navigate a web of shared meanings, and to make nuanced predictions about how others would do the same. As much as love and compassion, this is part of what it is to be human.

Acknowledgments

We thank Josh Tenenbaum, Laura Schulz, Steve Piantadosi, Shimon Ullman, Drazen Prelec, and the anonymous reviewers for their helpful comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2018.05.007.

References

Anderson, J. R. (2000). *Learning and memory: An integrated approach* (2nd ed.). New York: Wiley.
Bhatia, S. (2017a). Associative judgment and vector space semantics. *Psychological Review, 124*(1), 1–20.
Bhatia, S. (2017b). The semantic representation of prejudice and stereotypes. *Cognition, 164*, 46–60.
Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. *WW Norton & Company*.
Camerer, C. F., Ho, T.-H., & Chong, J. K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics,* 861–898.
Camerer, C. F., Ho, T.-H., & Chong, J. K. (2015). A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences, 3*, 157–162.
Carlin, G. (1972). Seven words you can never say on television. *Class Clown [CD]*. Santa Monica, California: Little David/Atlantic.
Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences, 3*(2), 57–65.
Christian, B. (2011). *The most human human.* New York: Anchor20.
Collins, H., & Evans, R. (2014). Quantifying the tacit: The imitation game and social fluency. *Sociology, 48*(1), 3–19.
Collins, H., Evans, R., Weinel, M., Lyttleton-Smith, J., Bartlett, A., & Hall, M. (2017). The Imitation Game and the nature of mixed methods. *Journal of Mixed Methods Research, 11*(4), 510–527.
Correll, J., Judd, C. M., Park, B., & Wittenbrink, B. (2010). Measuring prejudice, stereotypes and discrimination. *The Sage handbook of prejudice, stereotyping and discrimination* (pp. 45–62). London: SAGE.
Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*(4), 631.
der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(85), 2579–2605.
Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5.
Dovidio, J. F. (2010). *The SAGE handbook of prejudice, stereotyping and discrimination.* Sage Publications.
Downey, R. (Vol. Ed.), (2014). *Turing's Legacy: Developments from Turing's ideas in logic. Vol. 42*Cambridge University Press.
Epley, N., & Waytz, A. (2009). Mind perception. *Handbook of social psychology*.
Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878.
Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*(6084), 998.
French, R. M. (2000). The Turing Test: The first 50 years. *Trends in Cognitive Sciences, 4*(3), 115–122.
Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*(11), 818–829.
Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science, 5*, 173–184.
Gray, H., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619.
Gray, K., Jenkins, A. C., Heberlein, A. S., & Wegner, D. M. (2011). Distortions of mind perception in psychopathology. *Proceedings of the National Academy of Sciences, 108*(2), 477–479.
Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125–130.
Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4.
Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109*(1), 3.
Grice, H. (1975). *Logic and conversation*.
Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211.
Hilton, J. L., & Von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology, 47*(1), 237–271.
Jordan, M., & Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.
Kao, J. T., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. *Proceedings of the 36th annual meeting of the Cognitive Science Society* (pp. 719–724). .
Kao, J. T., & Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.
Kao, J. T., Levy, R., & Goodman, N. D. (2016). A computational model of linguistic humor in puns. *Cognitive Science, 40*(5), 1270–1285.
Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences, 111*, 12002–12007.
Klein, O., & Azzi, A. E. (2001). The strategic confirmation of meta-stereotypes: How group members attempt to tailor an out-group's representation of themselves. *British Journal of Social Psychology, 40*(2), 279–293.
Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*(12), 463–470.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.
Lewis, D. (1969). *Convention: A philosophical study*. John Wiley & Sons.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* (pp. 3111–3119). .
Miller, W. I. (1998). *The anatomy of disgust.* Harvard University Press.
Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402–407.
Operario, D., & Fiske, S. T. (2001). Stereotypes: Content, structures, processes, and context. *Blackwell handbook of social psychology: Intergroup processes. 1. Blackwell handbook of social psychology: Intergroup processes* (pp. 22–44).
Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP). 12. Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).
Pinker, S. (2007). *The stuff of thought: Language as a window into human nature.* Penguin.
Shieber, S. (1994). Lessons from a restricted Turing test. *Communications of the Association for Computing Machinery, 37*(6), 70–78.
Stahl, D., & Wilson, P. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior, 10*(1), 218–254.
Stuhlmüller, A., & Goodman, N. D. (2014). Reasoning about reasoning by nested

[10] While shibboleths may rely on shared culture and background to distinguish group members from outsiders, historically they often depended on pronunciation, as in the origin of the term: “The Gileadites captured the fords of the Jordan leading to Ephraim, and whenever a survivor of Ephraim said, “Let me cross over,’ the men of Gilead asked him, ‘Are you an Ephraimite?’ If he replied, No,” they said, “All right, say ‘Shibboleth.’” If he said, “Sibboleth,” because he could not pronounce the word correctly, they seized him and killed him at the fords of the Jordan.” (Judges 12:5–6, New International Version).

conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research, 28*, 80–99.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.
Ullman, T. D., Xu, Y., & Goodman, N. D. (2016). The pragmatics of spatial language. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.
Vorauer, J. D., Main, K. J., & O'Connell, G. B. (1998). How do individuals expect to be viewed by members of lower status groups? Content and implications of meta-stereotypes. *Journal of Personality and Social Psychology, 75*(4), 917.
Waytz, A., Gray, K., Epley, N., & Wegner, D. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14*(8), 383–388.
Wegner, D. M., & Gray, K. (2016). *The mind club: Who thinks, what feels, and why it matters.* Viking Adult.
Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.