Available online at www.sciencedirect.com

ScienceDirect

KELLEY SCHOOL OF BUSINESS
INDIANA UNIVERSITY
www.journals.elsevier.com/business-horizons

In bot we trust: A new methodology of chatbot performance measures

Aleksandra Przegalinska [a,*,^], Leon Ciechanowski [a,b,^],
Anna Stroz [c], Peter Gloor [d], Grzegorz Mazurek [a]

[a] Kozminski University, Jagiellonska 57/59, Warsaw, Poland
[b] University of Social Sciences & Humanities, Chodakowska 19/31, Warsaw, Poland
[c] University of Warsaw, Krakowskie Przedmieście 26/28, Warsaw, Poland
[d] MIT Center for Collective Intelligence, 245 First Street, E94-1509 Cambridge, MA, U.S.A.

Abstract Chatbots are used frequently in business to facilitate various processes, particularly those related to customer service and personalization. In this article, we propose novel methods of tracking human-chatbot interactions and measuring chatbot performance that take into consideration ethical concerns, particularly trust. Our proposed methodology links neuroscientific methods, text mining, and machine learning. We argue that trust is the focal point of successful human-chatbot interaction and assess how trust as a relevant category is being redefined with the advent of deep learning supported chatbots. We propose a novel method of analyzing the content of messages produced in human-chatbot interactions, using the Condor Tribefinder system we developed for text mining that is based on a machine learning classification engine. Our results will help build better social bots for interaction in business or commercial environments.
© 2019 Kelley School of Business, Indiana University. Published by Elsevier Inc. All rights reserved.

1. Human-machine interaction today

Artificial intelligence is an established and internally diverse academic discipline spanning many decades. It embraces such subfields as natural language processing, machine learning, robotics, and many others. With the major advancements in machine learning and the arrival of big data, it has become a discipline of utmost importance for the global economy and society as whole, as well as rapidly changing businesses and organizations. Current machine learning capabilities include competing at the highest level in strategic

---

* Corresponding author
    E-mail address: aprzegalinska@kozminski.edu.pl (A. Przegalinska)
    ^ A. Przegalinska and L. Ciechanowski contributed to this article equally.

games, autonomously operated vehicles, intelligent routing in content delivery networks, and—most importantly in the context of this article—understanding and generating natural speech. As we look at how far artificial intelligence is reaching, ask this question: How far can and should we reach with further developments in the responsiveness of chatbot systems?

Several real-world cases have shown the tremendous power that AI represents for companies. Facebook uses AI for facial recognition to make its applications better than humans at determining if two pictures are of the same person. Motorbike manufacturer Harley-Davidson increased New York sales leads almost 30-fold from less than five to around 40 per day through AI-enabled hyper-targeted and customized communications. These examples show that artificial intelligence, in addition to being a hot topic and a buzzword, has made its way into a variety of organizations and has become top-of-mind for many corporations and institutions. However, many issues are still unsolved. The more AI enters the world of business, the more technological, societal, and ethical questions arise.

In this article, we focus on the important chatbot subdiscipline within AI. Chatbots are interactive, virtual agents that engage in verbal interactions with humans. This technology is an interesting case in human-machine interaction as they are designed to interact with users through the usage of natural language based on formal models. This 60-year-old technology was initially designed to determine whether chatbot systems could fool users into believing that they were real humans. Quite early, it was confirmed that, when it comes to simulating interactions, cheating humans was possible (Khan & Das, 2018; Sloman, 2014). Recently, however, the main goal of building chatbot systems is not only to mimic human conversation and cheat users in various iterations of the Turing test (Aron, 2015; French, 2012; Hernandez-Orallo, 2000; Moor, 2003) but also to apply well-trained chatbots widely to interact with the users in business, education, or information retrieval. For quite a long time, applying chatbots was cumbersome and risky (Klüwer, 2011; Shawar & Atwell, 2007a). Currently, however, chatbots can facilitate various business processes, particularly those related to customer service and personalization because of their accessibility, fairly low cost, and ease of use for the end consumers. Our main goal in this article is to track the evolution of chatbots from simple syntactic systems to robust natural language processing entities that use deep learning (i.e., Google Duplex, Microsoft Cortana, GPT-2) and expose how this change effects business and organizations.

We propose a novel methodology of tracking human-chatbot interactions and measuring chatbot performance that takes into consideration ethical concerns—in particular, trust toward chatbots. Our research links neuroscientific methods, text mining, and deep learning with the issue of trust allocated in chatbots and their overall transparency. On a more general level, we address a wide range of issues of both theoretical and practical importance including the general effects of AI on companies and institutions, how AI should be used in order to create value, the potential negative effects of AI and how to overcome them, and ethical questions that emerged from the arrival of artificial intelligence.

2. Chatbots in organizations

One of the very first chatbots, ELIZA, was created in 1964—1966 at the MIT Artificial Intelligence Laboratory and was one of the first attempts to simulate human conversation. ELIZA (Weizenbaum, 1966; Weizenbaum & McCarthy, 1976) simulated conversations by using both a substitution approach and pattern matching that beguiled users with an illusion of understanding and reciprocity. Interaction directives hidden in scripts allowed ELIZA to process user logs and engage in communication. ELIZA's most known script was DOCTOR, which simulated a Rogerian psychotherapist representing an approach that empowers the patient in the therapeutic process. ELIZA creator Joseph Weizenbaum regarded the program as a method to expose the superficiality of human-machine communication but was surprised by the number of patients who attributed human-like feelings to ELIZA and wanted to continue their therapy with it. ELIZA's success was so surprising because the chatbot was incapable of understanding anything. While ELIZA could engage in discourse perceived by the interlocutor as fruitful, it could not converse on any deeper, semantic level, which led to a new idea for developing chatbots. Namely, many chatbot developers felt encouraged to use diverse types of tricks to deceive chatbot users. Instead of focusing on cumbersome work with natural language processing, they focused on the easiest possible ways to fool users. Such attempts were often successful as in the cases of PARRY (Curry & O'Shea, 2012; Phrasee, 2016) and

Eugene Gootsman (Aamoth, 2014; Shah, Warwick, Vallverdú, & Wu, 2016). Whereas the PARRY effectively simulated a paranoid person by iterating threats and suspicions in every conversation, Eugene Gootsman simulated a 14-year-old boy from Ukraine with limited knowledge of the world and a lower competence in speaking English.

We think of these scripts and default responses as the simulation paradigm. It evokes the Chinese room argument. However, with the rise of deep natural language processing capacities, this may soon change as more 'aware' systems will be created. Even before the arrival of graphical user interfaces, the 1970s and 1980s saw rapid growth in text and natural language interface research. Since then, a range of new chatbot architectures appeared, including MegaHAL, Converse, and A.L.I.C.E. (Shawar & Atwell, 2005, 2007b), which stores knowledge about the English conversation patterns in artificial intelligence markup language (AIML) files.

The choice of chatbots is much broader and includes a great deal of machine learning-supported consumer technologies like Siri and Cortana (Dunn, 2016; Hachman, 2014; López, Quesada, & Guerrero, 2017). These may assume the shape of virtual agents or physical objects which we could also research from the perspective of proxemic relations that they maintain with users, as well as gestural communication (Ciechanowski, Przegalinska, Magnuski, & Gloor, 2019).

Amazon launched the Alexa (Hayes, 2017) project as a platform for a better jukebox but the company developed it into an artificial intelligence system built upon and continuously learning from human data. The Alexa-powered cylinder Echo and the smaller Dot are omnipresent household helpers that can turn off the lights, tell jokes, or let users consume the news hands-free. They also collect reams of data about their users, which Amazon utilizes to improve Alexa and develop its possibilities (Hayes, 2017). Alexa maintains 15,000 skills but 14,997 of them remain untapped as most of the users are not aware of them (Przegalinska, 2018). Even though social robots like Alexa have progressed beyond ELIZA as far as usages and capacities in real-life, noisy environments, they are still embedded in the simulation paradigm. For this reason, both Google and OpenAI with their respective deep learning supported bots Duplex and GPT-2 offer Amazon hefty competition despite only boasting 378 skills for the Google Assistant (Leviathan & Matias, 2018), considering it has been training its AI with petabytes of data.

3. Current chatbot performance measurements

When it comes to performance measurements, there are certain standards even though the usage of metrics differs by industry (Botanalytics, 2018). Surveys of companies deploying chatbots (Asher, 2017; Castro et al., 2018) showed how companies across industries have been using chatbot analytics and which parameters they consider crucial in measuring performance. For instance, when it comes to banking and fintech sectors, chatbots are used mainly to assist users in performing tasks quicker, while lowering call volumes and cutting costs on the service (Okuda & Shoda, 2018). One of the most important chatbot performance metrics is conversation length and structure. Most of the industry trends suggest aiming for shorter conversations with simple structure because banking chatbots intend to provide fast solutions (e.g., sending/receiving money, checking a balance).

Retention rate is yet another important metric for measuring chatbot success. If a company's chatbot aims to replace other communication channels (e.g., lowering call volume), the goal is to obtain significantly higher retention, which is an indicator of consumer satisfaction. However, there are plenty of other automated options that allow users to manage accounts easily without speaking to a human. Thus, if the chatbot is focused more on customer support, a high retention rate does not necessarily have to be the measure of success.

Another measurement of chatbot performance is the ability to provide personalized communication. A financial chatbot should retain and analyze user information and subsequently provide users with relevant offers. For instance, Erica, a chatbot created for Bank of America, provides users with tips on how they can better save money depending on their spending habits.[1]

Retail chatbots provide gamified ways to shop. Through a conversational interface, brands transform themselves into personal shopping assistants. Unlike banking chatbots, retail chatbot creators usually look for a high number of conversation steps in chatbot engagement metrics. While users might seek a fast solution, retail chatbots are supposed to capture and hold users' attention by encouraging them to browse, providing responses about products or cross/upselling a purchase. An interesting example here is eBay's ShopBot, which

---

[1] https://promo.bankofamerica.com/erica/

offers several options to encourage browsing and purchasing.

4. Trust in the organizational context

Trust is generally defined as a firm belief in the reliability, truth, or ability of someone or something, or "an arrangement whereby a person (a trustee) holds property as its nominal owner for the good of one or more beneficiaries" (HOMS Solicitors, 2019). Trust among humans may be perceived in terms of credibility and confidence in one another's judgment, as well as predictability of one's behavior (Cassell & Bickmore, 2000).

In the organizational context, trust can be defined in multiple ways that possess a common ground of belief (Dunn & Schweitzer, 2005). It can be understood as:

- Competence, defined as "the belief that an organization has the ability to do what it says it will do…[including] the extent to which we see an organization as being effective; that it can compete and survive in the marketplace" (Institute for Public Relations, 2003);

- Integrity, defined as the belief that an organization is fair and just; or

- Reliability, defined as the belief that an organization "will do what it says it will do, that it acts consistently and dependably" (Grunig, 2002).

Although the experts are not in complete agreement, trust between an organization and its public is generally described as having various independently quantifiable characteristics including, but not limited to:

- Multilevel aspect—trust results from interactions within and between organizations;

- Communication and discourse-based aspect—trust emerges from interactions and communicative behaviors;

- Culture-driven aspect—trust is correlated with values, norms, and beliefs; and

- Temporal aspect—trust is constantly changing as it cycles through phases of establishing, breaking, discontinuing, etc. (Hurley, 2011; Shockley-Zalabak, Morreale, & Hackman, 2010).

In terms of linking trust and new technologies with AI in an organizational context, Kaplan and Haenlein (2019, p. 22) cited three common traits relevant "both internally and externally: confidence, change, and control—the three Cs of the organizational implications of AI." An important category we will elaborate on later in this article is external confidence, defined here as confidence in the abilities and recommendations of an organizations' AI systems.

5. Trust in chatbots

The issue of trust in bot performance is important for several reasons. Particularly, trust should be considered during the implementation of chatbots in environments such as financial care, healthcare, and other fields demanding sensitive data by which users may be exposed to physical, financial, or psychological harm (Bickmore & Cassell, 2001). Trust may be a crucial factor because users do not want to share personal information if they are unsure about the security (Chung, Joung, & Ko, 2017b).

A crucial part of trust is related to anthropomorphization (Ciechanowski et al., 2019; Lotze, 2016; Radziwill & Benton, 2017; Stojnić, 2015). Our intuition may lead us to claim that because chatbots interact by conversation, it is appropriate to anthropomorphize them. There are, however, two opposing positions. One of them claims that if an agent is more humanlike, then it is more probable that a sustainable trust relationship between agent and user will be made, known as a human-human perspective. Since humans place social expectations on computers, anthropomorphization increases users' trust in computer agents (i.e., the more human-like computer is, the more trust humans apply to it). The second point of view is that humans place more trust on computerized systems as opposed to other humans (Seeger & Heinzl, 2018). According to this view, high-quality automation leads to a more fruitful interaction because the machine seems to be more objective and rational than a human. Humans tend to trust computer systems more than other humans because humans are expected to be imperfect while the opposite is true for automation (Dijkstra, Liebrand, & Timminga, 1998; Seeger & Heinzl, 2018).

It is important to stress, however, that the process of anthropomorphization is not only about the attribution of superficial human characteristics but most importantly this essential one: a humanlike mind. People trust anthropomorphized

technology more than the mindless one, particularly because of the attribution of competence (Waytz, Heafner, & Epley, 2014) resulting in a significant increase of trust and overall performance of cooperation (de Visser et al., 2012; Waytz et al., 2014). The reliability of various elements of the digital economy ecosystem (e.g., wearable devices, systems and platforms, virtual assistants) arises as an important problem when one considers the level of trust allocated to them.

In a social context, trust has several connotations. Trust is characterized by the following aspects: one party (trustor) is willing to rely on the actions of another party (trustee), and the situation is directed to the future. In addition, the trustor (voluntarily or forcedly) abandons control over the actions performed by the trustee. As a consequence, the trustor is uncertain about the outcome of the other's actions; they can only develop and evaluate expectations. Thus, trust generally can be attributed to relationships between people. It can be demonstrated that humans have a natural disposition to trust and to judge trustworthiness that can be traced to the neurobiological structure and activity of a human brain. When it comes to the relationship between people and technology, the attribution of trust is a matter of dispute. The intentional stance demonstrates that trust can be validly attributed to human relationships with complex technologies and machine-learning-based trackers and sensors could be considered complex technologies (Taddeo, 2010). This is specifically true for information technology that dramatically alters causation in social systems: AI, wearable tech, bots, virtual assistants, and data. All that requires new definitions of trust.

In terms of trust in chatbots nowadays, one can distinguish at least two other relevant dimensions in addition to anthropomorphization (Nordheim, 2018):

- Ability/expertise: Performance measurements (e.g., customer retention, conversation length, lead generation). Expertise is seen as a factor associated with credibility, a cue for trustworthiness. In the context of an automated system, trust has been argued to be mainly based on users' perceptions of the systems' expertise. Here, the chatbot's expertise is assumed to impact users' trust.

- Privacy/safety: "Security diagnostics expose vulnerabilities and privacy threats that exist in commercial Intelligent Virtual Assistants (IVA)—diagnostics offer the possibility of securer IVA

ecosystems" (Chung, Iorga, Voas, & Lee, 2017a, p. 100).

In the past, one could argue the already mentioned Turing test was a specific form of trust-related practice. It was an experiment that was conceived as a way of determining the machine's ability to use natural language and indirectly prove its ability to think in a way similar to a human and, as such, it touched upon two dimensions of trust: anthropomorphism and expertise. In the case of the latter, trust was related to human expertise to properly assess the status of the interlocutor (human/nonhuman). Today, however, even transparency becomes a problematic topic. As mentioned before, Google Duplex in specific, narrowly-specialized topics exhibits a close-to-human level of human speech understanding and synthesis, as well ability to use nonverbal cues that the Turing test in previous forms would be inadequate for testing and analysis anymore. More importantly, in the samples of conversation provided by Google Duplex, testing did not reveal it was a bot and not a human being (Harwell, 2018). This is not to say that the Turing test should be suspended as a tool, but that it needs to be reformulated. What is more, DeepMind—a British AI company acquired by Google in 2014—is developing a project that may reshape the chatbot industry in an even more dramatic way.

DeepMind explores the field of the theory of mind (Premack & Woodruff, 1978) that broadly refers to humans' ability and capacity to observe various mental states as mental states belonging to other human beings. DeepMind proposed to train a machine to build such models and attempts to design a Theory of Mind[2] neural network (ToM-net) which uses metalearning to build models of the agents it encounters from observations of their behavior alone (Rabinowitz et al., 2018): "Through this process, it acquires a strong prior model for agents' behavior, as well as the ability to bootstrap to richer predictions about agents' characteristics and mental states using only a small number of behavioral observations."

---

[2] DeepMind applies the ToM-net to agents behaving in simple gridworld environments, showing that it learns to model random, algorithmic, and deep reinforcement learning agents from varied populations, and, more interestingly, passes classic ToM tasks such as the "Sally- Anne" test (Baron, 1985; Wimmer & Perner, 1983) of recognizing that others can hold false beliefs about the world. The authors argue that this system is an important step forward for developing multi-agent AI systems, for building intermediating technology for machine-human interaction, and for advancing the progress on interpretable AI.

Figure 1.   Two types of chatbots*
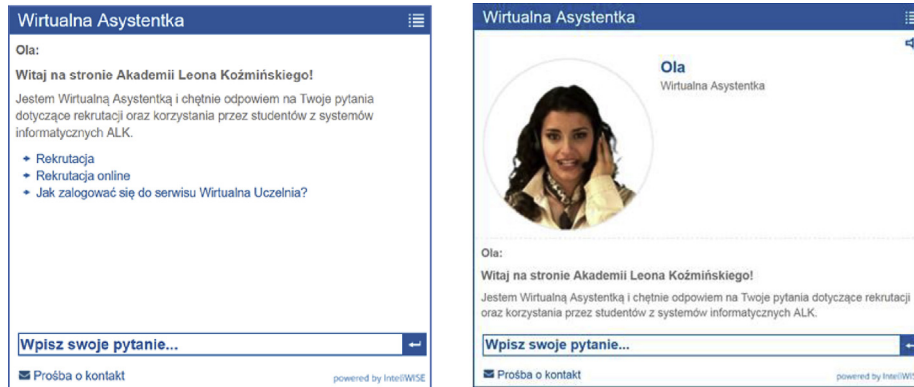
Note: Each user was randomly assigned to talk with one of them. On the left is the simple text chatbot, on the right is the avatar with sound.

What this implies in the context of our research—and more broadly—is that human-chatbot interaction is a new method of conceptualizing and researching trust. As we already discussed, in the case of nonblackbox systems (e.g., chatbots based on stochastic engines), attributing and researching trust was already difficult. With deep learning systems, new dimensions of trust-based human-machine relations open up.

6. Methodology

We conducted a study (the first of its kind) of human-chatbot online interaction (Ciechanowski et al., 2019; Ciechanowski, Przegalinska, & Wegner, 2018) using both subjective (questionnaires) and objective (psychophysiology) measures. The study consisted of interactions with one of two chatbot types: a simple text chatbot or an avatar with sound (Figure 1).

During the conversation with the chatbot, we gathered psychophysiological data from the participants and at the end of the procedure, they were asked to fill out a couple of questionnaires (Fiske, Cuddy, Glick, & Xu, 2002; Fong, Nourbakhsh, & Dautenhahn, 2002; Pochwatko et al., 2015) that took into account cooperation between humans and chatbots, realness of the chatbot, and its human likeness. The results important for the current study indicated that users were more satisfied and happier during the interaction with the simple text chatbot, which was confirmed in both questionnaire results and psychophysiological markers. These results have also proven the uncanny valley hypothesis (Mori,

MacDorman, & Kageki, 2012), which states that users and consumers are less satisfied when interacting with artificial objects ineptly imitating a human, and this is what our avatar chatbot was attempting to do. This result was also manifested directly in the uncanny valley measure that we applied in our experimental procedure.

However, the research question that was still interesting for us was related to linguistic correlations of the uncanny valley effect. In other words, perhaps the inept imitation of the avatar chatbot led the users to verbally behave in a specific way toward it, and the chatbot, therefore, responded in different ways to them, irritating them more than the simple text chatbot. In order to answer this question, we analyzed the messages users produced in their interaction with the avatar chatbot.[3]

For the analysis of the translated messages, we used Tribefinder, a novel machine-learning-based system able to reveal Twitter users' tribal affiliations (De Oliveira & Gloor, 2018; Gloor, Colladon, de Oliveira, & Rovelli, 2018). In this case, a tribe is "a network of heterogeneous persons…who are linked by a shared passion or emotion" (Cova & Cova, 2002, p. 602). In the market domain, users

---

[3] There are no tools for automatic text analysis or text mining in Polish; therefore, we needed first to translate the messages to English. For the purpose of translation of files with chatbot logs—users' prompts and chatbot's responses—from Polish to English language, the DeepL platform was used. DeepL Translator applies neural machine translation, resulting in impressive accuracy and quality of the text. Parts of sentences were translated manually—particularly because of some mental shortcuts and colloquial phrases introduced by users. The total number of translated entries was 1,373 for 13 users.

Figure 2. Mean line of user-chatbot exchange and number of messages for each user
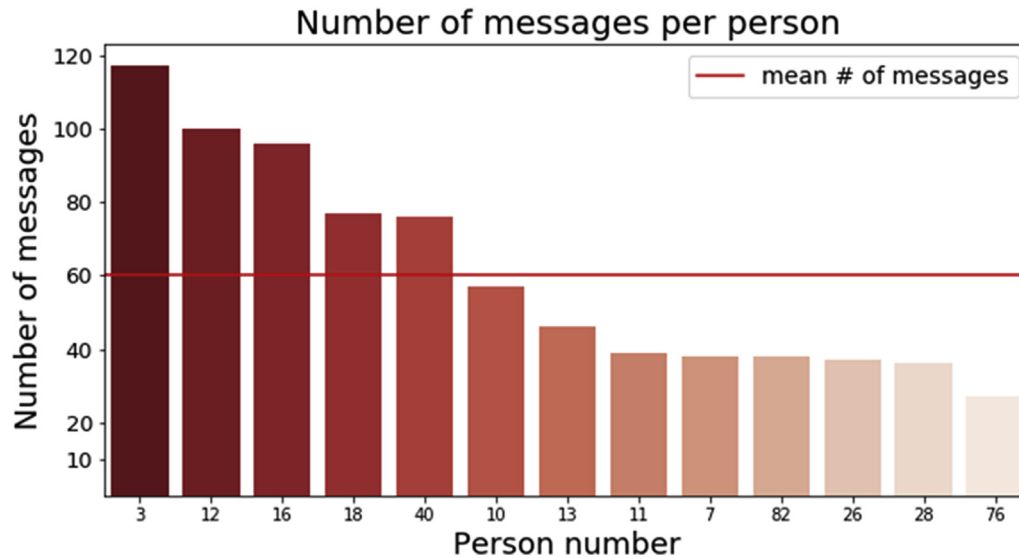
or consumers affiliate with different tribes and thereby can be characterized by various behaviors, rituals, traditions, values, beliefs, hierarchy, and vocabulary. The analysis and successful use of tribalism may be a crucial factor in modern marketing and management (Cova & Cova, 2002; Moutinho, Dionísio, & Leal, 2007).

Initially, the Tribecreator was used to select tribal influencers and leaders, and Tribefinder was created to establish tribes to which individuals belong through the analysis of their tweets and the comparison of vocabulary. Technically, Tribefinder is backed up by a machine learning system, which was trained with Twitter feeds of 1,000 journalists from *The New York Times*, *The Washington Post*, *The Guardian*, and with members of the U.S. Congress. We applied this system, in our case, to analyze the content of users' and chatbots' messages. Since we were interested in the reason for the users being less willing to cooperate with the avatar chatbot, we analyzed the interactions in this group using the Tribefinder.

The system was able to process messages from 13 out of 15 persons from the avatar chatbot group since it is based on machine learning algorithms and needed sufficient amounts of data in order to carry out a meaningful classification. The mean number of interactions (defined by the user-chatbot exchange of one message) of the selected users was 60.3 (Figure 2).

The Tribefinder system was able to allocate users into five tribe macro-categories (see Table 1):

- Personality;

- Ideology;

- Lifestyles;

- Alternative reality; and

- Recreation.

Ideology describes proponents of the main political ideologies, trained with 100 human representatives of each of the three main ideologies: socialism, capitalism, and liberalism. The creators of Tribefinder had to add the category of complainers to this list, as some people did not fit into any mainstream ideology (Gloor et al., 2018). Examples of messages based on which the Condor system placed bots and people into different categories can be found in the Appendix.

7. Results

After the analysis of messages generated by users and by the chatbot in the interaction with each user, we observed that all subjects and chatbots (except in two instances) were of the journalist personality type; therefore, they did not use deceitful language. Similarly, all persons and chatbots were complainers (except one person and one chatbot classified as liberal), and thus it was impossible to discern their ideology.

We observed that almost 50% of the users can be characterized as displaying characteristics of the

Table 1. Definitions of tribes assigned to users and chatbots by the Tribefinder system

Tribal macro-category	Tribe type	Characterization
Personality	Journalist	People stating the truth.
	Politician	People trying to lie or using deceitful language.
Ideology	Liberalism	People subscribing to the liberal ideology.
	Capitalism	People subscribing to the capitalist ideology.
	Socialism	People subscribing to the socialist ideology.
	Complainers	People not expressing any ideas related to the 3 main political ideologies.
Lifestyle	Fitness	People fond of doing sports, even to the point of being compulsive.
	Sedentary	The opposite of the fitness category, these people apply an inactive lifestyle.
	Vegan	People who avoid animal products in their diet.
	YOLO	People living according to the motto You Only Live Once. Can display hedonistic, impulsive, or reckless behavior.
Alternative reality	Fatherlander (Nationalist)	Conservatists, extreme patriots, family people, and xenophobes.
	Nerd (Technocrats)	Technology enthusiasts, transhumanists, fans of globalization and networking.
	Spiritualist	People believing in some sort of sacred power and meaning. Keen of spiritual contemplation.
	Treehugger (Environmentalists)	People stressing the importance of the protection of nature. Therefore, they accept some technologies (e.g., alternative energies), but challenge other (e.g., gene manipulation).
Recreation	Art	People interested in art, cherishing its beauty and emotions.
	Fashion	People following fashion (in the domain of clothing, hairstyle, or behavior).
	Sport	People following sport events on media. Some of them also actively practice sports.
	Travel	People loving to travel, getting to know different cultures and places.

Source: (Gloor et al., 2018)

vegan tribe, 33% as a YOLO, and 25% as a sedentary (Figure 3). We found that every instance of a chatbot interacting with these users was assigned by the Tribefinder to the sedentary tribe. The chatbot used in our experiment is stationary and nonrobotic; therefore, this affiliation is correct.

When it comes to the alternative realities, we observed an interesting picture. Looking at the users, only two categories emerge: spiritualism (69%) and nerd (31%). However, when we looked at the tribes the chatbots affiliated with this category, we observed all four tribes represented (Figure 4), with the tribes of treehugger and nerd standing out. Interestingly enough, an opposite situation takes place with the recreation category (Figure 5).

These results indicate that users can be mostly described as using spiritual language in conversations with chatbots. This becomes more understandable when we combine it with the fact that the users knew that they were having a conversation with an artificial being and were provoking the chatbot to say something human. Moreover, the users agreeing to participate in our experiment were likely to be technology enthusiasts, thereby the other alternative reality tribal category they showed was nerd. The above-mentioned differences between humans and bots are more visible when presented on a stack column chart (see Figure 6).

Figure 6 renders the similarities and differences between human and bot messages. Evidently, the bot adjusted its messages to humans in the categories of personality and ideology, while there were significant differences in the categories of

Figure 3. Lifestyles of users
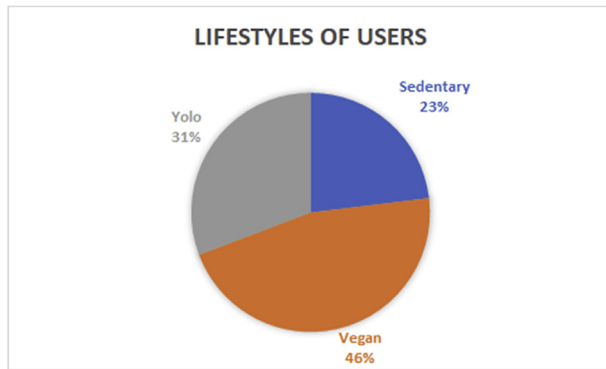
lifestyle, alternative realities, and recreation. These results are understandable as the chatbot has been following humans in the most abstract and general categories like using mostly truthful language (journalist type) and, just like human participants, the bot manifested no political beliefs (complainer type). This, however, may have serious consequences. Let us imagine a chatbot that represents a sports company versus clients. If the chatbot mirrors statements of a user that hates sports, it will say things that go against the company's mission, vision, and values. Such a chatbot will most definitely not be an interface that does its job well.

8. Essentials in chatbot measurements for emerging chatbots

Looking at the Gartner (2018) Hype Cycle, technologies that have not lived up to their hype end up in the obsolete before plateau category. Given the significant gap between Alexa skill development and use, Alexa and many other chatbots may be classified in this category. Several chatbots and

social robots may turn out to be useful only as an interface to a search engine—as the thing that asks follow-up questions to refine the user's search to find exactly what the user is looking for.

On the other hand, developers believe that voice-based AI devices are not just jukeboxes that iterate the simulation paradigm with better technologies. In fact, chatbots are the perfect example of the implementation of state-of-the-art consumer-oriented artificial intelligence that does not only simulate human behavior based on formal models but also adapts to it. As such, chatbots remain a fascinating subject for the research of patterns in human and nonhuman interaction along with issues related to assigning social roles to others, finding patterns of (un)successful interactions, and establishing social relationships and bonds. In that sense, chatbots are gradually becoming robust, context-aware systems. Let us not forget that scientists and engineers initially created chatbots because people wanted to use natural language to communicate with computer systems smoothly. (Zadrozny et al. 2000, pp. 116—117) argued that the best way to facilitate human-computer interaction is by allowing users "to express their interests, wishes, or queries directly and naturally, by speaking, typing, and pointing." Morrissey and Kirakowski (2013) made a similar point in their criteria for the development of a more human-like chatbot. Perhaps chatbots should become more sensitive to human needs by searching and delivering information that people need and would otherwise be unable to obtain (Kacprzyk & Zadrozny, 2010; Morrissey & Kirakowski, 2013).

Taking into account that scientists and researchers are developing bots and chatbots so that they become more refined, nuanced, context-aware, and transparent collaboration assistants, there are few relevant changes related to redefined dimensions of trust. Namely, on the basis of our research and the major developments in

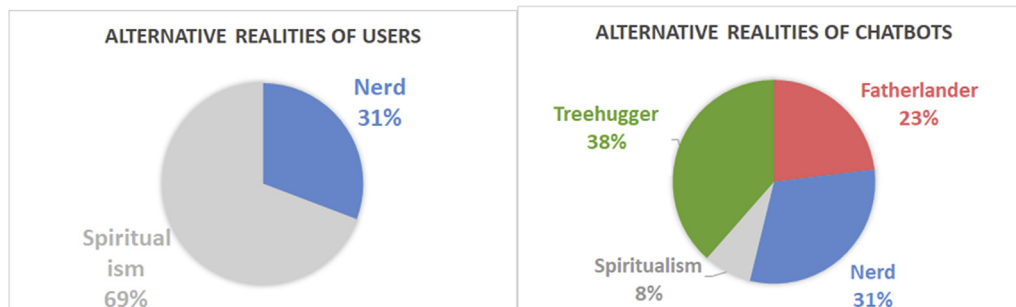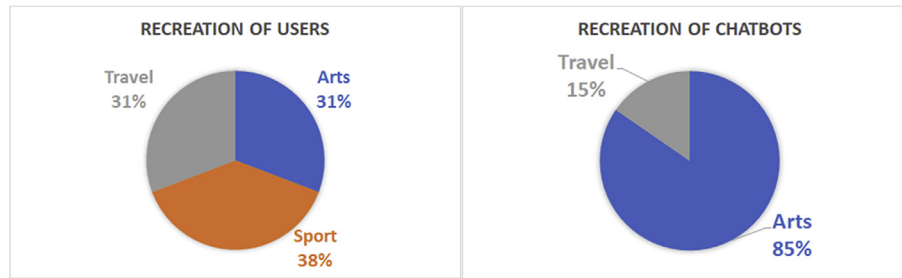Figure 4. Alternative realities of users and chatbots

Figure 5.   Recreation of users and chatbots

botics, there are new dimensions of trust we would like to suggest:

- Transparency and honesty: the agent sends honest signals in communication, does not speak like a politician (according to Tribefinder categories), and it does not deny its status.

- Predictability, already mentioned in publications concerning trust and chatbots (Nordheim, 2018), could be supplemented by integrity, seen as a factor associated with credibility, and concerns the trustors' expectation that an object of trust will act consistently in line with past experience. If the user perceives the

chatbots as predictable, this may lead to a feeling of trust in the chatbot.

- Control and benevolence describe the degree to which the motivations and intents of the trustee are in line with those of the trustor.

This last element is strongly linked with the challenge of unblackboxing AI and introducing the greater explainability of information processing performed by AI (Petkovic, Kobzik, & Re, 2018; Preece, 2018). Integrity and benevolence could be further researched by Tribefinder and similar tools, as well as applied in business chatbots.

Figure 6.   Human and chatbot tribe types
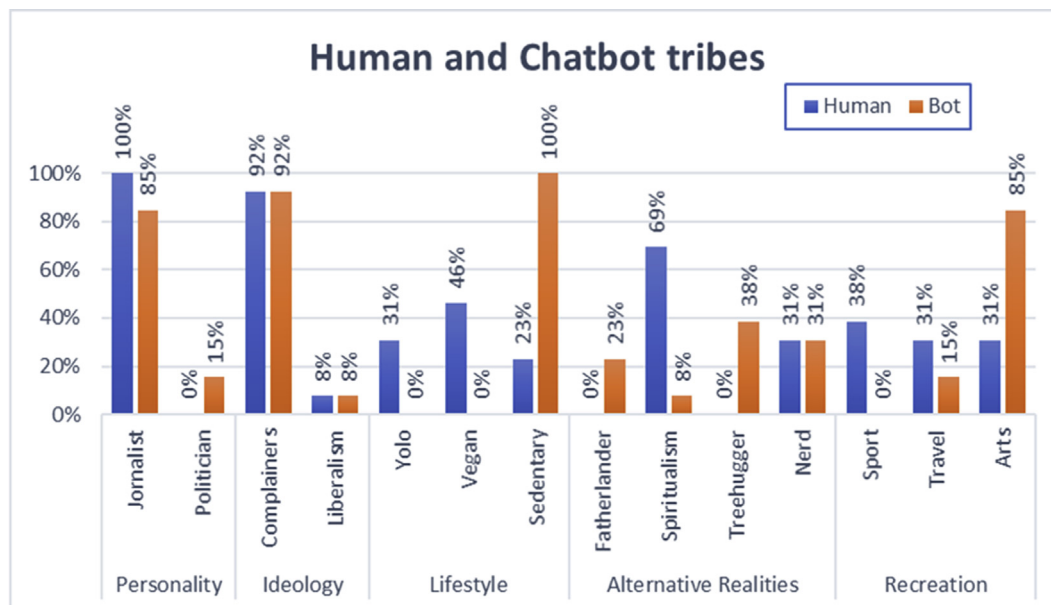
Note: Presented in separate tribal macro-categories (e.g., personality). Percentage shows how many persons/bots were in a specific tribal macro-category, for instance, in the personality tribal macro-category 100% humans were categorized as journalists, and—in consequence—0% of them were categorized as politicians.

9. Recommendations for business

Whether we want it or not, bots and chatbots are already becoming actors in the online world. Working on new methods of creating trustworthy conversational agents for the future is a task of science, research, and business with business ultimately implementing the methods. Taking into account the emerging trend of cobotization (Méda, 2017), finding paths toward creating safe, useful, and robust chatbot solutions is a very important task. Therefore, our recommendation is a focus on the three new trust-enhancing features of chatbots that we enumerated in the previous section (i.e., transparency, integrity, explainability) that allow for greater control. What remains to be seen is whether humanoid bots and robots—often employed in the sales or customer service departments—are truly able to interact with humans and how much real interaction and cooperation between a human being and a chatbot or robot is possible.

Appendix

References

Aamoth, D. (2014). Interview with Eugene Goostman, the fake kid who passed the Turing test. *Time*. Available at: https://time.com/2847900/eugene-goostman-turing-test/

Aron, J. (2015). Forget the Turing test—there are better ways of judging AI. *NewScientist*. Available at: https://institutions.newscientist.com/article/dn28206-forget-the-turing-test-there-are-better-ways-of-judging-ai/

Asher, N. (2017). *A warmer welcome: Application of a chatbot as a facilitator for new hires onboarding*. Available at: http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1116842&dswid=4708

Baron, J. (1985). *Rationality and intelligence*. Cambridge, UK: Cambridge University Press.

Bickmore, T., & Cassell, J. (2001). Relational agents: A model and implementation of building user trust. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 396—403). New York, NY: SIGCHI.

Botanalytics. (2018). *How chatbot performance metrics differ by industry*. Available at: https://chatbotslife.com/how-chatbot-performance-metrics-differ-by-industry-b380d4bd7f6b

Cassell, J., & Bickmore, T. (2000). External manifestations of trustworthiness in the interface. *Communications of the ACM, 43*(12), 50—56.

Castro, F., Tedesco, P., Alves, H., Quintino, J. P., Steffen, J., Oliveira, F., et al. (2018). Developing a corporate chatbot for a customer engagement program: A roadmap. In *Intelligent computing theories and application* (pp. 400—412). Cham, Switzerland: Springer International.

Chung, H., Iorga, M., Voas, J., & Lee, S. (2017a). Alexa, can I trust you? *Computer, 50*(9), 100—104.

User message	Chatbot message
What did you do today?	I think I'm working around the clock.
Tired?	Please try to formulate your question differently.
Are you tired?	I am not tired, rather the opposite. I am full of energy.
That's good	Great!
Great	I'm glad, too.
What are your plans for the evening?	I will sit here. That's how I spend most of the evenings.

This excerpt is from the user nr 3. Both the user and this chatbot instance were assigned by the Condor system to the journalist personality type (using truthful language).

User message	Chatbot message
Hi :)	Hey!
I am Oskar	I am very pleased about this!
How are you?	Great, a lot of clients.
Nice weather.	I haven't left since yesterday, don't ask me about the weather.
Ok, I have few questions.	Every day I have a different color - 365 types for the whole year.
Aha, great.	I'm glad.
I have few questions.	I'm waiting for your questions.

This excerpt is from the user nr 16. The user was assigned by the Condor system to the journalist personality type, while this chatbot instance assigned to the politician (using deceitful language).

Chung, M., Joung, H., & Ko, E. (2017b). The role of luxury brand's conversational agents: Comparison between face-to-face and chatbot. In *Proceedings of the 2017 global fashion management Conference at Vienna*. Available at: http://gfmcproceedings.net/html/sub3_01.html?code=326161

Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2019). The shades of the uncanny valley: An experimental study of human—chatbot interaction. *FGCS: Future Generation Computer Systems, 92*, 539—548.

Ciechanowski, L., Przegalinska, A., & Wegner, K. (2018). The necessity of new paradigms in measuring human-chatbot interaction. In M. Hoffman (Ed.), *Advances in cross-cultural decision making*. Cham, Switzerland: Springer International.

Cova, B., & Cova, V. (2002). Tribal marketing: The tribalisation of society and its impact on the conduct of marketing. *European Journal of Marketing, 36*(5/6), 595—620.

Curry, C., & O'Shea, J. D. (2012). The implementation of a story-telling chatbot. *Advances in Smart Systems Research, 1*(1), 45—52.

De Oliveira, J. M., & Gloor, P. A. (2018). GalaxyScope: Finding the "truth of tribes" on social media. In F. Grippa, J. Leitão, J. Gluesing, K. Riopelle, & P. Gloor (Eds.), *Collaborative innovation networks: Building adaptive and resilient organizations* (pp. 153—164). Cham, Switzerland: Springer International.

de Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., et al. (2012). The world is not enough: Trust in cognitive agents. In *Proceedings of the annual meeting of the human factors and ergonomics society* (pp. 263—267). Santa Monica, CA: HFES.

Dijkstra, J. J., Liebrand, W. B. G., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology, 17*(3), 155—163.

Dunn, J. (2016). We put Siri, Alexa, Google Assistant, and Cortana through a marathon of tests to see who's winning the virtual assistant race — here's what we found. *Business Insider*. Available at: https://www.businessinsider.com/siri-vs-google-assistant-cortana-alexa-2016-11

Dunn, J. R., & Schweitzer, M. E. (2005). Feeling and believing: The influence of emotion on trust. *Journal of Personality and Social Psychology, 88*(5), 736—748.

Fiske, S. T., Cuddy, A. J. C., 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—902.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2002). *A survey of socially interactive robots: Concepts, design, and applications (CMU-RI-TR02-29)*. Pittsburgh, PA: Carnegie Mellon University.

French, R. M. (2012). Moving beyond the turing test. *Communications of the ACM, 55*(12), 74—77.

Gartner. (2018). *5 trends emerge in the Gartner Hype Cycle for emerging technologies, 2018*. Available at: https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018/

Gloor, P. A., Colladon, A. F., de Oliveira, J. M., & Rovelli, P. (2018). Identifying tribes on Twitter through shared context. In Y. Song, F. Grippa, P. A. Gloor, & J. Leitão (Eds.), *Collaborative innovation networks*. Cham, Switzerland: Springer.

Grunig, J. E. (2002). *Qualitative methods for assessing relationships: Between organizations and publics*. Available at: https://instituteforpr.org/wp-content/uploads/2002_AssessingRelations.pdf

Hachman, M. (2014). Battle of the digital assistants: Windows phone Cortana vs Google now vs Siri. *PC World*. Available at: https://www.pcworld.com/article/2142022/the-battle-of-the-digital-assistants-windows-phone-cortana-vs-google-now-vs-siri.html

Harwell, D. (2018). A Google program can pass as a human on the phone. Should it be required to tell people it's a machine? *The Washington Post*. Available at: https://www.washingtonpost.com/news/the-switch/wp/2018/05/08/a-google-program-can-pass-as-a-human-on-the-phone-should-it-be-required-to-tell-people-its-a-machine/?utm_term=.cf13615d7898

Hayes, A. (2017). *Amazon Alexa: A quick-start beginner's guide*. Scotts Valley, CA: CreateSpace Independent Publishing Platform.

Hernandez-Orallo, J. (2000). Beyond the turing test. *Journal of Logic, Language and Information, 9*(4), 447—466.

HOMS Solicitors. (2019). *Trusts: Tax implications*. Available at: https://www.homs.ie/publications/trusts-tax-implications/

Hurley, R. F. (2011). *The decision to trust: How leaders create high-trust organizations*. Hoboken, NJ: John Wiley & Sons.

Institute for Public Relations. (2003). *Guidelines for measuring trust in organizations*. Available at: https://instituteforpr.org/guidelines-for-measuring-trust-in-organizations-2/

Kacprzyk, J., & Zadrozny, S. (2010). Computing with words is an implementable paradigm: Fuzzy queries, linguistic data summaries, and natural-language generation. *IEEE Transactions on Fuzzy Systems, 18*(3), 461—472.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15—25.

Khan, R., & Das, A. (2018). Introduction to chatbots. In R. Khan, & A. Das (Eds.), *Build better chatbots: A complete guide to getting started with chatbots* (pp. 1—11). Berkeley, CA: Apress.

Klüwer, T. (2011). From chatbots to dialog systems. In D. Perez-Marin, & I. Pascual-Nieto (Eds.), *Conversational agents and natural language interaction: Techniques and effective practices* (pp. 1—22). Hershey, PA: IGI Global.

Leviathan, Y., & Matias, Y. (2018). Google Duplex: An AI system for accomplishing real-world tasks over the phone. *Google AI blog*. Available at: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

López, G., Quesada, L., & Guerrero, L. A. (2017). Alexa vs. Siri vs. Cortana vs. Google assistant: A comparison of speech-based natural user interfaces. In I. L. Nunes (Ed.), *Advances in human factors and systems interaction* (pp. 241—250). Cham, Switzerland: Springer.

Lotze, N. (2016). Chatbots. *Peter Lang*. Available at: https://www.peterlang.com/view/title/18967

Méda, D. (2017). The future of work: The meaning and value of work in Europe. *HAL Archives*. Available at: https://hal.archives-ouvertes.fr/hal-01616579/

Moor, J. H. (2003). The status and future of the Turing Test. *Minds and Machines, 11*(1), 77—93.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley. *IEEE Robotics and Automation Magazine, 19*(2), 98—100.

Morrissey, K., & Kirakowski, J. (2013). 'Realness' in chatbots: Establishing quantifiable criteria. In M. Kurosu (Ed.), *Human-computer interaction: Interaction modalities and techniques* (pp. 87—96). Berlin, Germany: Springer.

Moutinho, L., Dionísio, P., & Leal, C. (2007). Surf tribal behaviour: A sports marketing application. *Marketing Intelligence & Planning, 25*(7), 668—690.

Nordheim, C. B. (2018). *Trust in chatbots for customer service: Findings from a questionnaire study*. Available at: https://www.duo.uio.no/bitstream/handle/10852/63498/1/CecilieBertinussenNordheim_masteroppgaveV18.pdf

Okuda, T., & Shoda, S. (2018). AI-based chatbot service for financial industry. *Fujitsu Scientific and Technical Journal, 54*(2), 4—8.

Petkovic, D., Kobzik, L., & Re, C. (2018). Machine learning and deep analytics for biocomputing: Call for better explainability. *Pacific Symposium on Biocomputing, 23*, 623—627.

Phrasee. (2016). *PARRY: The AI chatbot from 1972*. Available at: https://phrasee.co/parry-the-a-i-chatterbot-from-1972/

Pochwatko, G., Giger, J.-C., Różańska-Walczuk, M., Świdrak, J., Kukiełka, K., Możaryn, J., et al. (2015). Polish version of the negative attitude toward robots scale (NARS-PL). *Journal of Automation Mobile Robotics and Intelligent Systems, 9*(3), 65—72.

Preece, A. (2018). Asking 'why' in AI: Explainability of intelligent systems — perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management, 25*(2), 63—72.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515—526.

Przegalinska, A. (2018). *Wearable technologies in organizations: Privacy, efficiency and autonomy in work*. New York, NY: Springer.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Ali Eslami, S. M., & Botvinick, M. (2018). Machine theory of mind. *Cornell University*. Available at: https://arxiv.org/abs/1802.07740.pdf

Radziwill, N. M., & Benton, M. C. (2017). *Evaluating quality of chatbots and intelligent conversational agents*. Cornell University. Available at: http://arxiv.org/abs/1704.04579.pdf

Seeger, A.-M., & Heinzl, A. (2018). Human versus machine: Contingency factors of anthropomorphism as a trust-inducing design strategy for conversational agents. In F. Davis, R. Riedl, J. vom Brocke, P.-M. Léger, & A. Randolph (Eds.), *Information systems and neuroscience* (pp. 129—139). Cham: Springer International.

Shah, H., Warwick, K., Vallverdú, J., & Wu, D. (2016). Can machines talk? Comparison of Eliza with modern dialogue systems. *Computers in Human Behavior, 58*, 278—295.

Shawar, B. A., & Atwell, E. S. (2005). Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics, 10*(4), 489—516.

Shawar, B. A., & Atwell, E. (2007a). Chatbots: Are they really useful? *LDV Forum, 22*(1), 29—49.

Shawar, B. A., & Atwell, E. (2007b). Different measurements metrics to evaluate a chatbot system. In *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies* (pp. 89—96). Stroudsburg, PA: Association for Computational Linguistics.

Shockley-Zalabak, P. S., Morreale, S., & Hackman, M. (2010). *Building the high-trust organization: Strategies for supporting five key dimensions of trust*. Hoboken, NJ: John Wiley & Sons.

Sloman, A. (2014). *Judging chatbots at turing test 2014*. Available at: https://www.cs.bham.ac.uk/research/projects/cogaff/misc/turing-test-2014.pdf

Stojnić, A. (2015). Digital anthropomorphism: Performers, avatars, and chat-bots. *Performance Research, 20*(2), 70—77.

Taddeo, M. (2010). Trust in technology: A distinctive and a problematic relation. *Knowledge, Technology & Policy, 23*(3), 283—286.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113—117.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36—45.

Weizenbaum, J., & McCarthy, J. (1976). *Computer power and human reason: From judgment to calculation*. New York, NY: W.H. Freeman and Company.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103—128.

Zadrozny, W., Budzikowska, M., Chai, J., Kambhatla, N., Levesque, S., & Nicolov, N. (2000). Natural language dialogue for personalized interaction. *Communications of the ACM, 43*(8), 116—120.