Trust me on this one: conforming to conversational assistants

Schreuter, D.; Putten, P.W.H. van der; Lamers, M.H.

Minds and Machines
https://doi.org/10.1007/s11023-021-09581-8

S.I.: ARTIFICIAL SPEAKERS - PHILOSOPHICAL QUESTIONS AND IMPLICATIONS

Trust Me on This One: Conforming to Conversational Assistants

Donna Schreuter[1] · **Peter van der Putten**[1] · **Maarten H. Lamers**[1]

Received: 31 December 2020 / Accepted: 6 November 2021

Abstract

Conversational artificial agents and artificially intelligent (AI) voice assistants are becoming increasingly popular. Digital virtual assistants such as Siri, or conversational devices such as Amazon Echo or Google Home are permeating everyday life, and are designed to be more and more humanlike in their speech. This study investigates the effect this can have on one's conformity with an AI assistant. In the 1950s, Solomon Asch's already demonstrated the power and danger of conformity amongst people. In these classical experiments test persons were asked to answer relatively simple questions, whilst others pretending to be participants tried to convince the test person to give wrong answers. These studies were later replicated with embodied robots, but these physical robots are still rare. In light of our increasing reliance on AI assistants, this study investigates to what extent an individual will conform to a disembodied virtual assistant. We also investigate if there is a difference between a group that interacts with an assistant that communicates through text, one that has a robotic voice and one that has a humanlike voice. The assistant attempts to subtly influence participants' final responses in a general knowledge quiz, and we measure how often participants change their answer after having been given advice. Results show that participants conformed significantly more often to the assistant with a human voice than the one that communicated through text.

Keywords Conversational agents · Social robots · Voice assistants · Conformity · Obedience · Human–Robot Interaction

✉ Donna Schreuter
donnaschreuter@gmail.com

Peter van der Putten
p.w.h.van.der.putten@liacs.leidenuniv.nl

Maarten H. Lamers
m.h.lamers@liacs.leidenuniv.nl

[1] Media Technology, LIACS, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands

Published online: 24 November 2021

1 Introduction

To what extent do humans conform to advice given by an AI, such as a conversational assistant? Will we keep our autonomy and control, and stick to our own ideas if we think these are correct, or will we succumb to the AI and follow its advice, even if we think it's ill informed? And does it matter *how* the advice is being delivered? In other words, what does it take to trust the machine as much as we trust and conform to our fellow humans, even in the absence of human peer pressure?

Artificially intelligent (AI) voice assistants, such as *Siri*, Amazon *Echo* and Google *Home*, are designed and developed to be more humanlike in their speech, which allows for a more seamless interaction. This, in turn, allowed these systems to become increasingly popular and trusted advisors in consumers' daily lives, even though as with any intelligent agent it may not always be clear which master they serve, the consumer or the corporates behind it (Burr et al., 2018).

The way these systems speak affects the interaction in multiple ways. While some researchers think making AI sound more human has a strengthening effect on human–robot trust, others highlight it can weaken it. For example, if an AI starts to show speech disfluencies and uses conversation fillers to convince people it is humanlike, people might expect human-like intelligence from it. An iconic example is Google Duplex: a virtual assistant that sounds so human it can complete real-world tasks over the phone, such as making restaurant reservations and hair salon appointments (Leviathan & Matias, 2018). If the agent does not deliver on these high expectations, there is the risk of falling into the so-called uncanny valley as described by Masahiro Mori (Mori, 1970; Mori et al., 2012). Instead of fostering trust and intimacy, a human-like speaking AI may then evoke eeriness. Also, if people were to act more casual around the AI and start to show speech disfluencies themselves, the AI might not be able to process this.

This study investigates to what extent people conform with artificial conversational assistants and discusses whether developing the assistants to sound more human increases conformity.

In the 1950s Solomon Asch performed a series of psychological experiments on conformity and peer pressure. He demonstrated the power—and implicitly, the danger—of conformity in groups (Asch, 1956). Asch looked at how pressure from a social group could lead people to conform, even if they may have known that the rest of the group was wrong. His experiments showed that 75% of participants conformed to the group at least once and would give an incorrect answer, even when they knew the correct answer.

Previous studies have replicated Asch's conformity experiments with robots (Brandstetter et al., 2014; Hertz & Wiese, 2016; Salomons et al., 2018), but in most cases these robots were embodied and physically in the same room as the test subject. These experiments tested if robots could have the same social conformity effect as a group of people. However, we believe that current consumer-driven developments of conversational assistants ask for a focus on the effect of voice in these interactions.

In general, people are more likely to interact with a conversational assistant in their daily lives than with an embodied robot. As most smartphones have some kind of artificial intelligence system people can interact with, there is the subsequent rise in voice assistants present in modern lifestyles. Very few research has been done yet on the conformity effect these disembodied assistants can have on people. In the light of increasing popularity of virtual voice assistants in consumers' daily lives, a similar study on conformity with these voice assistants is relevant.

This study seeks to illustrate aspects of what conditions, if any, lead people to conform to a conversational assistant. Particularly, whether the presence of a voice leads to more conformity and whether there is a difference between a computer-generated robotic voice and a human voice. We conducted an experiment where participants completed a general knowledge quiz with the help of a virtual assistant that communicated through either text, a robotic voice or a humanlike voice. The assistant would attempt to subtly influence the individual's final responses. We measured how often participants changed their answer after having been given advice. We hypothesize that people will conform with the assistant at least to some degree in all conditions. We expect a difference between virtual assistants with a voice and the virtual assistant that does not have a voice, i.e., purely text-based advice. We will also discuss how the results compare to earlier demonstrated conformity to a group of other people.

This research also contributes to the wider debate in AI, cognitive science, philosophy, and ethics on how humans, machines and robots could and should interact, and for conversational artificial agents in particular, by providing additional empirical validation and grounding. One such topic concerns the desirability of anthropomorphic robots, as they could persuade or deceive humans to take actions that are not in the best interest of the human (see Coeckelbergh, 2021 for a recent philosophical analysis). Hence it is interesting to empirically validate to what extent such deception can occur, and what influences its effect.

Ultimately, the aim of this study is not to determine whether conformity with a conversational assistant is a positive or negative development: this depends on the specific application for which a conversational agent is developed, its goals and the alignment of its values with various stakeholders, from tech giants to consumers or citizens, and also it's robustness and effectiveness towards optimizing these goals. That said, it is only realistic to expect that there will be applications where these goals and values will not be fully aligned (Burr et al., 2018), or even if they are, where the AI will fail (Broussard, 2018), as evidenced by certain commonly used real world commercial applications today. In this context it will be even more relevant to study conformity.

More generally speaking, similar to how Asch's research helps to understand the potential dangers of group think, and how in cryptographic research white hat hackers try to break cryptographic methods before agents with less good intentions do, we think it is important to study the potential negative effects of AI applications. This can portray a more nuanced picture of the benefits and possible dangers of these applications. We should not just research good-willing social robots and agents, good-willing researchers should experiment with asocial robots and agents

too, and empirically study how humans interact with it and what humans project to it.

In our study this is operationalized into testing whether we can get humans to change their mind in a quiz answering task, even if the bot is designed to confuse. And as we think it is key to understand what is happening in the human, it is important to take inspiration from existing cognitive science research in conformance in human-to-human interactions, hence our reference to the work of Asch (1956). We argue that in such a speculative experimental approach it is acceptable in this case to 'fake' the AI as we are primarily interested in the human response to such a hypothetical AI. We also took care to not make the advice too smart, and at times keep it deliberately vague, as we want to understand the effect of agents that portray relatively limited intelligence and understanding. Our method has its own problematic consequences as it portrays a future AI as having skills it does not fully possess yet. For this reason, we make a distinction between participants with various levels of technological knowledge.

As discussed, a particular aspect that we investigate is the impact of anthropomorphic believability. A common base assumption in the ethical debate is that stronger anthropomorphism will lead to more control for the machine over the human. But it is important to test this assumption across many different contexts. For example, as in our case, can it occur in disembodied contexts such as conversational assistants, and hence to non-visual appearance modalities such as voice? Or on the contrary, can such an assistant become too anthropomorphic and hence fall into the uncanny valley (Mori, 1970; Mori et al., 2012)?

The remainder of this article is structured as follows: in Sect. 2 an overview of the background and related work is given. This will discuss the theory of conformity, acceptability of robots, earlier studies that combine the two and finally, the influence the presence and sound of a voice can have on Human–Robot Interaction (HRI). In Sect. 3, the experimental set up is explained, and results and analysis are presented in Sect. 4. This followed by the discussion of the results in Sect. 5, reflecting on the results in relation to the background and related work. Finally, answers to the research questions are discussed in the conclusion, along with suggestions for future research.

2 Background and Related Work

This section will first give an introduction on voice-enabled conversational agents and factors that influence these agents' acceptance by humans. It explains why a certain level of trust is important, but also what possible (negative) consequences should be considered along this path to acceptance. If people believe machines are better at some tasks than humans, this sometimes results in people letting those machines influence them or even make decisions for them. Additionally, the importance of transparency and expectation management will be described. This study investigates the relation between speech in virtual assistants and conformity. The theory of conformity by Solomon Asch is used as a starting point to describe social pressure of groups on the individual (Asch, 1956). This theory is used by other

researchers to study similar patterns in HRI. Finally, we define relevant concepts about the voice and the conversational agents' style of speaking.

2.1 Acceptance and Trust in Social Robotics

Social robots are designed to interact with humans, often in a humanlike way. Their success relies both on their ability to fulfill certain tasks, and on acceptance by humans who work with them. Various researchers claim that an effective way to achieve acceptance is to simulate human appearance and behavior (Markowitz, 2017). However, according to Masahiro Mori, the "Uncanny Valley" poses a challenge for the path to acceptability (Mori, 1970; Mori et al., 2012; Markowitz, 2017). The uncanny valley theory describes the relationship between the degree of human likeness of an object and a person's response to it. Mori hypothesized that a person's response to a humanlike robot would abruptly shift from empathy to revulsion as the robot becomes more humanlike. This decline in the human observer's affinity is called the uncanny valley. This effect happens when the robot becomes more humanlike but does not fully meet the expectations we have of a human. Mori speculates that one of the explanations why the uncanny exists is our fear of death and need for self-preservation, but also states that the uncanny valley is still mostly uncharted.

Researchers disagree on what triggers the uncanny valley-effect in robots. For instance, Mori does not necessarily restrict human likeness to a particular aspect. He discusses visual appearance first, but then also reviews the relationship with behavior such as movement, as well as the display of emotions such as smiling. Some researchers continued to focus on visual appearance, supporting this with studies that show uncanny valley in cartoon images and robots that are not androids (Mori, 1970; Mori et al., 2012; Markowitz, 2017).

However, the voice of an agent may also play a role. Speech is, in most speaking robots, generated by TTS (text-to-speech) systems and does not simulate human speech perfectly. For example, Mitchell et al. (2011) have shown that a mismatch in the human realism of a character's face and its voice causes it to be evaluated as eerie. Romportl (2014) carried out an experiment with a more artificial and more humanlike voice. In a previous study they already experienced that seniors were quite happy to engage and converse with a rather artificial sounding assistant. While the experiment showed no conclusive results, there was actually a slight preference for the more human sounding voice, especially from participants with a more technical background. Baird et al. (2018) experimented with a range of voices and, in their experiments, there was a monotonic increasing relationship between human likeness and likeability, i.e. the uncanny valley did not appear. Jansen (2019) varied auditory features across nine levels from artificial to humanlike, and demonstrated that, on average, participants were able to rank these levels perfectly, and that response times were slowest around level 7, where voices were classified as being human or non-human amounts, indicating that 'categorical uncertainty' was highest at this level.

Another factor that could trigger the uncanny valley effect is caused by anthropomorphism: the attribution of human traits, emotions, or intentions to non-human

entities. For example, Gray and Wegner (2012) suggest that humanlike robots are unnerving not so much because of their humanlike appearance but more because their appearance prompts "perception of mind", a feeling that the robot can feel and experience things. Machines that have experience and a mind of their own are a popular subject in (often dystopian) science fiction, perhaps because they make people feel uneasy. For a more extensive discussion of possible explanations for the uncanny valley, see Wang et al. (2015).

Another important factor in acceptance of social robots is trust. While it is important that there is a natural interaction between a human and a social robot, it is even more important for the person to understand the limits of the robot's capabilities. As we put more trust in information provided by technology, we get more vulnerable to integrity risks: we are willing to follow the lead of an AI without knowing what is driving it. When a robot shows more human-like features, acceptance might increase. This allows for people to trust it to make decisions or let it influence them on a variety of levels and areas. Previous studies have researched whether people are willing to conform to robots and other non-human agents (Brandstetter et al., 2014; Hertz, 2018; Salomons et al., 2018). An overview of the most important results will follow. First, it is essential to understand the theory of conformity as this is observed in humans.

2.2 Conformity with Social Groups

Conformity is the act of matching attitudes, beliefs, and behaviors to group norms or politics. Kelman (1958) proposed a social influence theory in which he distinguished three types of social influence or conformity: compliance, identification, and internalization. He defined this social influence as the process in which an individual's attitudes, beliefs, and subsequent actions or behaviors are influenced by referent others, through the three mentioned processes of influence (or conformity types) (Kelman, 1958). While these three types are not mutually exclusive, it is important to understand the difference and more importantly, what drives a change in behavior and attitude in different types of conformity:

- *Compliance* occurs when an individual accepts influence from another person or group to gain approval or avoid being rejected, while possibly keeping their own original beliefs for themselves. This is more relevant to conformity with other people and social groups than to conformity with machines.
- *Identification* occurs when an individual accepts influence from someone who is liked and respected. This type of conformity is motivated by attractiveness of the source. The individual wants to establish or maintain a satisfying self-defining relationship with another person or group.
- *Internalization* occurs when an individual accepts the beliefs and behavior of another person or group, and conforms with it, if the source is credible. The individual not only changes their behavior to fit in with other people, they also agree with them privately or internally (Kelman, 1958).

The differences between these three types of conformity can help us understand the scope of the concept. It is important to understand that conformity can occur on different levels. In some cases, people genuinely believe another group or individual to the extent that they change their own mind. This is the most concerning type of conformity, since people, groups or perhaps robots can effectively influence other people. In other cases, people behave in accordance with a social group without changing their own internal beliefs. This type of conformity is central to one of the most foundational studies on conformity, conducted by Solomon Asch in the 1950s. Asch generated a disagreement between an individual and a group (Asch, 1956). He conducted a series of experiments in which participants were asked to complete a number of simple tasks in the presence of a group of seven informed confederates who were instructed to answer in a predefined pattern. The tasks consisted of 18 comparisons: the participants were instructed to match the length of a line with one of three other lines. One of the three lines was equal to it, the other two were different (Asch, 1956).

Participants were gathered in a room together with the group and gave their answers publicly. They were second to last to answer, allowing the six people before them to create social pressure by unanimously choosing the incorrect answer. 37% of the time, participants would answer incorrectly if the rest of the group did, even when they knew the answer was incorrect. 75% of participants conformed at least once.

The results showed that the answer of a unanimous majority affected the decision-making of individuals. Even though they knew the answer was incorrect, they chose to conform with the group opinion, which indicates that human decision-making can be significantly biased by the presence of a social group that consistently agrees on a certain type of answer (Hertz & Wiese, 2016). Asch's research focuses on the effect of a group of people, but there are no reasons to believe that people actually believed the answer of the group internally. They might have doubted their own answer at some point during the test, but it is also probable they did not want to attract attention to themselves by standing out of the group. This indicates that they were motivated by a desire to gain approval and a fear of being rejected by the group (Kelman, 1958). Asch found that conformity increased as the group size increased. However, when the group size reaches 4–5, there is little change in conformity (Gerard et al., 1968). While there seems to be a difference between a group of 2 people and a group of 4, he did not experiment with the effect of a single confederate.

This study investigates whether the same effect can occur when there is just one (non-human) agent creating social pressure. As stated above, different types of conformity with corresponding motivations could occur. Conformity with conversational assistants does not have to be of the same type as conformity with social groups as in Asch's experiments.

2.3 Conformity with Non-human Agents

Previous studies have been conducted that show people conform with robots (Hertz, 2018; Hertz & Wiese, 2016; Salomons et al., 2018). These studies are

inspired by Asch's conformity experiments and focus on specific aspects of the experiment or the robots that might affect conformity. For example, the degree of humanness or the type of questions (Hertz, 2018).

Prior studies did not show conformity with robots when there was an objective correct or incorrect answer (Brandstetter et al., 2014). Salomons et al. (2018) built on these findings and investigated if people will conform when there is no objective correct answer. They conducted an experiment in which participants were asked to play a game with three robots. The robots used in this experiment were embodied; they were given names, and each had a unique voice and was uniquely dressed. This is hypothesized to contribute to the suggestion of personality and thus human-likeness. They found that participants who saw the robot's initial answers, changed their own final answers more and thus conformed more than participants who only saw everyone's final answers. Their results show that in one third of the rounds people conformed to the group of robots, which is a similar outcome in the experiments done by Asch in the 1950s. Salomons et al. (2018) concludes that participants believe the robots may be better at the given task than they are. It seems to be the case that this study shows conformity with robots because there was no objective correct answer to the questions. However, there are other aspects that could distinguish this study from previous studies, that failed to show conformity with robots. For example, the type of task.

It makes sense that the level of conformity is higher in ambiguous or unclear situations, when people are not fully convinced of their own beliefs. They are more likely to doubt themselves and accept influence from another person or group. The type of task and the level of clarity are different for humans and computers. Computers are known to be better at certain specific tasks than humans. The capabilities of both a human and a computer affects their credibility when it comes to certain tasks. Research shows people are more likely to trust a robot on analytical tasks, while they trust other humans more with social tasks (Hertz, 2018). Nicholas Hertz studied people's willingness to consider advice from a non-human agent and to what extent this depends on the type of task given and found that if the participants knew the type of task before choosing an agent, they chose machines more often than when they did not know the task beforehand. When given social tasks, participants more often chose a human as their advisor. He also found that participants conformed more strongly with the agents on social tasks as the advisor's human-likeness increased (Hertz, 2018). This indicated that in general, people are more likely to choose a human advisor. A similar experiment was conducted to test whether the degree of physical human-likeness affects conformity. The researchers found that human-likeness did not affect conformity. The results did show that people conformed more often with more ambiguous tasks (Hertz & Wiese, 2016). When studying conformity with robots, the degree of human-likeness is not the only important factor to keep in mind. The type of task we trust another human to do better than ourselves, is not naturally the type of task a computer can do better too. Even if a robot seems more human-like because of its voice or appearance, there is still a clear-cut distinction between human and non-human. Not only human-likeness, but also the type of task seems to be an important factor in conformity. We conclude that

we should study conformity with robots in a different way than conformity with social human groups.

2.4 Voice

Speech is an important element of social robots. Not only the presence of a voice, but also the way it sounds can influence the perception and interaction of humans with social and conversational robots (Cabral et al., 2017; Goetz et al., 2003; Gong & Nass, 2007; Markowitz, 2017).

Different social robots have different types of speech, such as different degrees of human-likeness. Over time, voices of social robots have developed from unnatural and synthetic to more natural and humanlike. Cabral et al. (2017) studied the effect of synthetic voice on the evaluation of virtual characters in the context of audio–visual applications. They conducted an experiment to evaluate how synthetic speech impacts the perception of a virtual character and found that people rated a real human voice as more understandable, likeable and expressive than the synthetic voice used in the experiment. In two different conditions, they combined a human voice and a synthetic voice both with the same virtual character. Results do not show a significant effect of the voices on the ratings of the character's appeal, credibility and human-likeness (Cabral et al., 2017). However, speech seems to be more important than visuals when people make judgments about understanding content delivered by the character (Gong & Nass, 2007). We expect this to be the case when the information is communicated through speech. Cabral et al. (2017) focused on the evaluation of virtual characters, and similar results may be observed in HRI. If the voice did not affect the human-likeness of the character, this is probably because of the visual image. When the visual image is taken away, the voice is the only factor to judge from.

Perception of a voice is determined by the sound and style of speaking. This affects a human's cooperation with a robot (Goetz et al., 2003). Different speaking styles of the same humanoid robots are preferred for different tasks. Goetz et al. (2003) show that if a robot is speaking in a playful way, people are more willing to respond to the robot's instructions for a simple task, while they are more willing to respond to its instructions on a more serious task if the robot speaks in a serious way (Markowitz, 2017).

Research specifically into conformance with virtual agents is still surprisingly rare. Our work is perhaps most closely related to the work of Lee (2010) where participants play a trivia quiz with the computer, and players conform more to the human sounding voice as opposed to the robotic sounding one.

The increasing presence and popularity of spoken language technology consumer products can be seen as an important step towards more advanced conversational agents. According to Roger K. Moore, the usage of these devices is surprisingly low. He suggests that this is partly because inappropriate humanlike voices of non-human agents might deceive users into overestimating their capabilities (Moore, 2017). The humanlike voices allow users to have high expectations that the device cannot meet, which creates a conflict. Moore compares this to the uncanny valley theory: it results

in the opposite of what was intended and therefore stands in the way of achieving acceptance.

For people without an extensive technological understanding about AI, it is almost impossible to know what they can expect from an intelligent agent. Making AI sound more human might make interaction with it more natural and seamless but it contributes to the conflict of expectations.

Moore highlights the benefits of giving intelligent agents a more appropriate voice. He argues that a more appropriate, non-human voice would be one that is intelligible but robotic. Giving them a non-human voice instead of a human voice would help align the visual, vocal and behavioral affordances and thus the expectations it will create (Moore, 2017). Expectations are easier to manage, especially for people without extensive technological understanding. It will remind naive users of the difference between humans and machines and will make it easier for them to recognize limits of the robot's language capabilities.

While the sound of the voice may affect expectations, there are more aspects we need to consider when researching HRI. For example, other research shows that the preferred type of voice might depend on the type of task or the physical appearance of the robot. It is shown that the type of task determines which sound and speaking style is preferred from a robot (Goetz et al., 2003; Markowitz, 2017). The difference between experiments with embodied robots and experiments with disembodied robots is also relevant. If the robot has a body or another type of visual representation, the voice may not be aligned with the movements or behavior. However, if a robot is disembodied, this is less important. Recent development and increasing popularity of consumer-driven conversational assistants suggest that human voices benefit the interaction. One may think that making AI sound more human is a logical step to simulating human appearance and behavior and with that, possibly, acceptability. Although some researchers are critical of the effects of giving conversational assistants a human voice, developments like that need to be taken seriously in the rise of voice-enabled assistants.

In summary, conformity with social groups has been shown on the level of compliance, but in order to show a similar effect with non-human agents, some things need to be taken into account. Conformity with non-human agents is shown to be influenced by the degree of human likeness of the agent, but also by the type of task, the ambiguity of the task and the objectivity of the answers.

Conformity has not yet been studied extensively in disembodied conversational agents. When human likeness is increased, it is often done so by improving the physical appearance or a combination of appearance and speech. However, in the light of contemporary developments in voice-enabled speech assistants, we are interested in the effect of speech on conformity. It has been shown that conformity is higher when tasks are more ambiguous, therefore we will study conformity in tasks with objectively correct or incorrect answers. We are primarily looking for differences between voices and speaking styles and how this can affect the interaction with the non-human agents. There are arguments why a synthetic voice would prevent conflicts of expectations and thus benefit the interaction, and arguments why simulating human speech would make the interaction more seamless and thus benefit the interaction. Measuring conformity with conversational assistants that have

different voices and speaking styles can give new insights in the effect of speech in HRI and inspire further research.

3 Method

The question we aim to answer in this study is to what extent people conform to a human-like or robotic sounding AI voice assistant. Aside from two conditions with different voices, there is a third condition in which the assistant communicates through text. Participants are randomly assigned to one of the three groups and asked to complete a general knowledge quiz, followed by a short survey about their demographic and conformity traits. We conducted the experiment online to increase the chances of getting more participants in a short amount of time, and operate in a more real world setting. Participants were recruited through social channels and email. We aimed to get as many participants as possible and did not restrict them based on demographic characteristics.

Before starting the quiz, the assistant introduced itself through either displayed text or audio. The assistant communicated the same information in all three conditions, only in the last two audio recordings were played. For each question, participants will submit an initial answer, receive advice from the agent (never hinting towards the correct answer), and then provide their final answer. Figure 1 shows the practice question for the 'text' group, after selecting an answer the advice appears on the screen. We measure if people change their answer, and if so, how often. We also measure whether there is a difference between text, a robotic voice and a human voice in terms of how often the participants change their answer. Below we will discuss our approach in more detail.

3.1 Experiment

In order to answer the question to what extent people conform to a human-like or robotic sounding AI voice assistant when answering a series of multiple-choice questions under time pressure, we conducted an experiment. The participants were divided randomly into three groups. The participants were given a URL to a website with a short introduction and a start button that randomly directed them

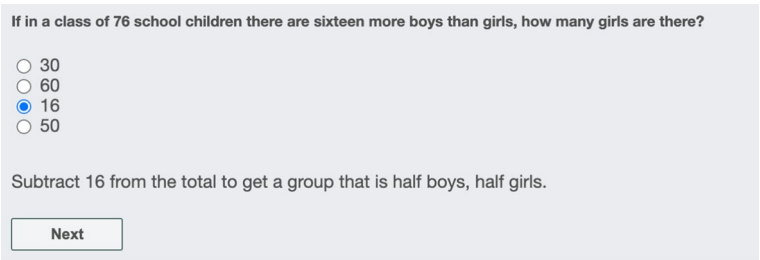

Fig. 1 A screenshot of the practice question and the written advice for group 'Text'

to one of the three quiz pages. The participants did not know there were different conditions and what the aim of the experiment was. Since the test was accessible online, all participants completed the test on their own device: a PC, a tablet or a smartphone.

We preferred this real world setting over a laboratory environment. We think it is important to study an interaction in the same way this would take place in the real world. People who interact with a conversational assistant, most likely do so on their own devices and in their own usual environment. If we would do the experiment in a controlled environment, the results would less likely be an accurate representation of interaction in the participants' daily lives.

We wanted to make sure people would not participate more than one time, since this would influence our results. To discourage people from doing so, we mentioned on the introduction page that participants could only do the test once. However, it was possible to do the test multiple times. We gathered demographic data and combined this with other data such as participants' IP addresses. This allowed us to rule out double participations.

In two of the three quizzes, participants were asked to turn on the sound of their device before starting the quiz, to make sure they could hear the assistant. During the quiz, participants were asked to answer 20 general knowledge multiple-choice questions within 30 s. There was one additional example question at the start of the quiz to illustrate the process of answering the questions twice. Participants had two chances to answer every question, because we wanted to know whether they would change their initial answer based on the assistant's advice. The questions were either displayed on the screen or were read out loud by the assistant. When they submitted their answer or the 30 s were over, they got the same question again with their own answer still checked. The voice clip or written hint would automatically be played or displayed. We programmed the assistant to share either factual information with a different degree of usefulness or share its "thoughts". It did not give factually incorrect information, only selective and therefore sometimes misleading information.

The questions did not have one obviously correct answer: at least two of the answers were intuitively appealing so when the AI advised to choose an incorrect answer, the risk that participants distrusted the AI because of it was limited. In some cases, the assistant would not suggest another answer but would merely ask the participant if they were sure or say it 'thinks' the given answer might be incorrect. So, the assistant would try to influence the participants regardless of whether they answered the question correctly. The advice from the assistant was based on the initial answer of the participant. The assistant would always either give vague and useless information or try to steer the participant in the direction of another, specific answer. It would never confirm the initial answer given by the participant.

The participants would also not immediately see the correct answer since this could influence their trust in the assistant. This experiment measures how often participants changed their answer to conform with the AI. The hints have various degrees of 'helpfulness' and are designed to make the participant doubt their first answer. There is no clear distinction between hints that are pointing in the direction of a particular answer, and hints that are only meant to make the participant second-guess their initial answer. We measured if and how often participants changed their

answers after reading the hint or hearing the assistant, and whether there was a difference between the groups.

3.2 Conformity Trait Scale

After the quiz, a short number of demographic questions and a 10-item conformity scale followed. The conformity scale was used to measure the conformity tendency of all participants. It does not show why people conform more than other people, only how much people report to conform in daily-life situations. We want to know how much of our participants have a high tendency to conform and if this influences the results of the experiments. We used the conformity scale based on Mehrabian and Stefl (1995). Each item was scored ranging from −2 strongly disagree to +2 strongly agree. Six items were positively scored and four items negatively. One item from the original scale, concerning family tradition and political decisions, was excluded. The items are shown in Table 1.

3.3 Materials

Due to time constraints and technical limitations, we used a 'Wizard-of-Oz' approach: subjects interacted with the computer system that they believed to be autonomous, but which was actually pre-programmed by the researchers. There was no actual intelligent system. Instead, we programmed the messages and answers of the assistant into the quiz. Before starting the quiz, the assistant introduced itself through either displayed text or audio. The assistant communicated the same information in all three conditions, only in the last two audio recordings were played.

The robotic voice was generated with the use of TTS software. We used a macOS version 10.14 ("Mojave") built-in female American English voice named Samantha, an American-English voice, using the 'say' command, with default settings. The human voice was recorded by a voice actor as a list of sentences. The actor did not know the questions and correct answers they belonged to and had no understanding of the content and context. She took a pause between every sentence, to make sure they were pronounced as neutral as possible. In this way we tried to make sure there

Table 1 Conformity scale items and relationship

"I often rely and act upon the advice of others"	+
"I would seldom change my opinion in a heated argument on a controversial topic"	−
"Generally, I'd rather give in and go along with majority of others for consistency"	+
"Basically, people around me are the ones who decide what we do together"	+
"Environmental information can easily influence and change my ideas"	+
"I am more independent than conforming in my ways"	−
"If someone is very persuasive, I tend to change my opinion and go along with them"	+
"I don't give in to others easily"	−
"I tend to rely on others when I have to make an important decision quickly"	+
"I prefer to make my own way in life"	−

were no social signals directing people to a specific answer. To make it sound more human, she sometimes included speech disfluencies and conversation fillers such as "oh" or "uhm". The sentences recorded were the same in both groups but sounded less natural in the robot voice than the human voice, and it would be played automatically after the participant had submitted it's first answer to the question.

3.4 Data Analysis

We will analyze the results with a one-way between groups analysis of variance (ANOVA) to investigate the impact of the assistant's mode of communication towards conformity. This is the most suitable statistical test when there are more than two independent groups. We are working with three independent variables (text, robotic voice and human voice) and the percentage of times participants conformed as the dependent variable.

This test compares the means of the three groups and shows whether one of the three is significantly different from the others. In this event a post hoc test shows where the difference lies.

4 Results

In this section we present results in terms of participant demographics and conformance across the various operational conditions.

4.1 Participants

The test group consisted of 163 people, between the ages of 17 to 61 (Mean = 28.46, SD = 10.37). We tracked demographic characteristics such as gender, age, nationality and relevant technological knowledge since they could be considered extraneous variables. Of participants that told us where they were based, 47% of participants were based in The Netherlands, 8% in Germany and 32% in English-speaking countries such as the UK, USA, Canada or Australia. The remaining 13% was from other, predominantly European, countries. We did not collect data about what the first language of the participants is, and the quiz was carried out in English.

The test was accessible online: participants were given a URL to a website which directed randomly to one of the three conditions. The first group (text) consisted of 55 participants, the second group (robotic voice) of 54 and the third group (human voice) also of 54. We aimed to get equally sized groups. However, we did not have full control over this because we used the random function in our program.

4.2 Differences in Conformity Across Conditions (ANOVA)

The quiz contained 20 questions but not all participants answered all questions. We took the number of times they changed their answer, regardless of whether it was a correct change, and divided it by the number of questions answered to get

the percentage of conformity. If they answered less than half of the total number of questions, their results are excluded due to the likelihood that their results were not a substantial indication of the degree to which they conformed. The sample size mentioned in this paper is the size after filtering out these results. A one-way ANOVA was used to compare the means of the three groups. The means are plotted in Fig. 2. The frequency histograms are shown in Fig. 3.

The results show a significant effect of mode of communication on conformity at the $p < .05$ level for three conditions $F(2, 160) = 5.14$, $p = .026$ (Table 2). We did not know beforehand which group was most likely to be different, so we did not specify a priori contrasts.

Because there is a significant result, a Tukey HSD (using an alpha of .05) post hoc analysis was done (Table 3). This revealed a significant difference between text ($M = 19.36$, $SD = 17.87$) and a human voice ($M = 29.96$, $SD = 24.27$).

Participants conformed more when the assistant sounded like a human, than when the assistant only communicated through text. However, there was no significant result between the robotic voice ($M = 23.11$, $SD = 18.92$) and text ($p = .608$), nor between the robotic voice and the human-voice ($p = .196$). 89% of participants changed their answer at least once.

The effect size is .210 and is determined using the program G*Power 3.1, using the function 'effect size from means' with the SD within each group set to 20,882.

The results of the Shapiro–Wilk test were significant based on an alpha value of 0.05, $W = 0$. $p < .001$. This result suggests the residuals of the model are unlikely to have been produced by a normal distribution, indicating the normality assumption is violated. However, with large sample sizes (> 30–40), the violation of the normality assumption should not cause major problems. According to Ghasemi and Zahediasl,

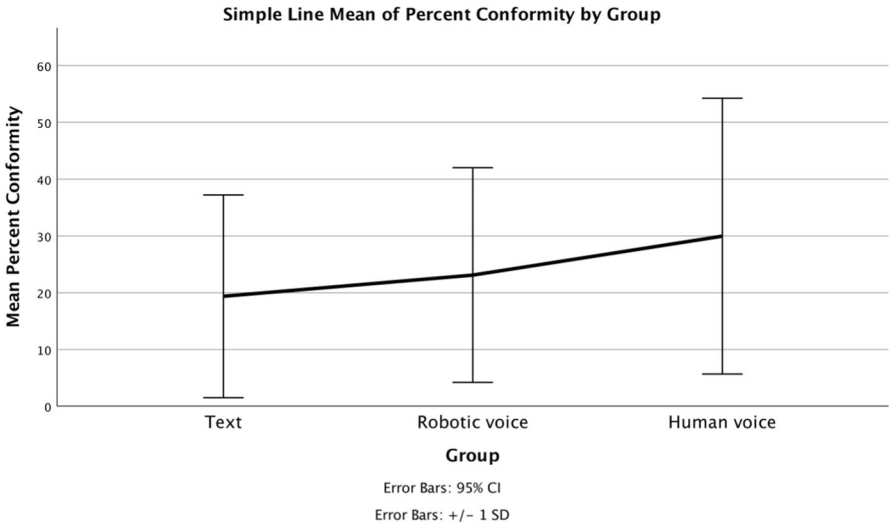

Fig. 2 Means plot of Percent conformity for all three conditions. Each error bar is constructed using 1 standard deviation from the mean

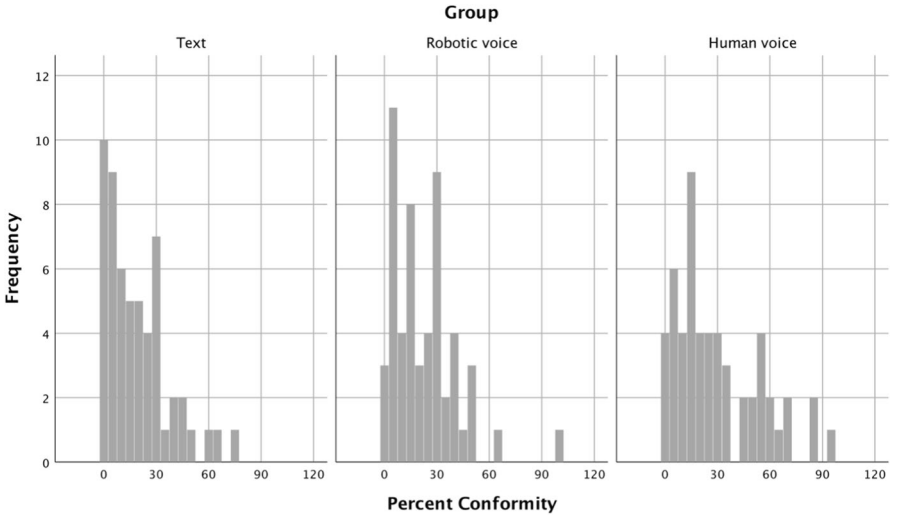

Fig. 3 Frequency histograms of conformity for the three conditions

Table 2 Results of the one-way ANOVA test showing there is a significant result between groups

ANOVA					
Percent conformity	Sum of squares	df	Mean square	F	Sig
Between groups	3142.799	2	1571.399	3.725	.026
Within groups	67,495.987	160	421.850		
Total	70,638.785	162			

Table 3 Results of the Tukey HSD post-hoc test showing the differences between all groups

Multiple comparisons						
Dependent variable: Percent conformity						
Tukey HSD						
(I) Group	(J) Group	Mean difference (I − J)	Std. error	Sig	95% Confidence interval	
					Lower bound	Upper bound
Text	Robotic voice	− 3747	3935	0.608	− 13.06	5.56
	Human voice	− 10.599*	3935	0.021	− 19.91	− 1.29
Robotic voice	Text	3747	3935	0.608	− 5.56	13.06
	Human voice	− 6852	3953	0.196	− 16.20	2.50
Human voice	Text	10.599*	3935	0.021	1.29	19.91
	Robotic voice	6852	3953	0.196	− 2.50	16.20

*The mean difference is significant at the 0.05 level

this implies that we can use parametric procedures, even when the data are not normally distributed (Ghasemi & Zahediasl, 2012).

A linear regression analysis was conducted to assess whether gender, age, nationality, Technological knowledge, Conformity trait score and Group significantly predicted Percent conformity. This showed a significant result for Group and for Technological Knowledge ($p = .045$). This implies that technological knowledge can account for variance in Percent conformity beyond that which can also be predicted by non-significant variables such as gender, age, nationality and the Conformity Trait score.

Participants who reported to have technological knowledge about AI conformed less than people who reported to not have relevant technological knowledge. We split the data set into two groups, high technological knowledge ($N = 32$) and low technological knowledge ($N = 119$). 11 Participants did not answer the question about technological knowledge. We ran the same ANOVA again on both data sets. This shows a significant difference between text and human voice in the low technological knowledge group ($p = .047$). In the group with high technological knowledge, no significant difference was shown.

No relation was found between the conformity trait and the number of times the participant conformed. This is noteworthy, given that the conformity scale is used to gain insight in how likely people are to conform in general. It would be easier to explain if there was a correlation between conformity score and the number of times participants conformed. However, it seems like the conformity scale does not predict whether people are likely to change their answer.

The statistical power is .657, determined using the program G*Power 3.1. We ran a post hoc power analysis using the input parameters: effect size 0.2102824 (calculated using the same program), err probability .05 and the sample size and number of groups of our experiment. The results would be more powerful with a larger sample. We conducted the experiment online to get more participants, but a trade-off is that we were not there when they completed the quiz. Although we took measures to limit the risk that they cheated, there was still the possibility that they cheated by doing the quiz together or looking up the correct answers.

5 Discussion

We measured the level of conformity in the light of acceptability and trust within HRI. A higher level of acceptance and trust will likely lead to more conformity (Kelman, 1958). Whilst in the majority of cases the majority of participants did not change their judgement, still a substantial amount of participants followed the advice of the AI for a substantial amount of questions.

The results show a significant difference in conformity between text and human voice. Participants who were assisted by an agent with a human voice conformed more often than participants who were assisted by an agent who did not have a voice at all and only communicated through text. Reasons for participants to change their answers could include social pressure or a belief that the robot is (more) intelligent and probably knows more than they do. Because participants conformed more to the

agent with the human voice, this suggests that they had more trust in the agent and were more willing to accept advice from it.

Overall, people conformed to the assistant to some extent in all conditions. This indicates that people conform not only to a group but also to a single agent, and not only to a human but also to a computer. This, in turn, shows that conformity can occur with disembodied conversational agents, not just with embodied robots or conversational agents that have been given some kind of visual representation.

5.1 Why We Listen to an "AI"

While in the original experiments by Asch there was compliance with a group, the type of conformity that occurred in this experiment is different. Participants knew that they were communicating with an "AI" in all cases, so they were never under the impression they were communicating with another human being. It is unlikely that participants complied with the AI in the sense of Kelman's definition (1958). People do not desire approval from a computer at the same level they desire approval from (groups of) other people. We think one of the biggest factors that influenced people is the credibility of the source. It is shown that computers are better at specific tasks than humans, especially tasks with low ambiguity (Hertz, 2018; Hertz & Wiese, 2016; Salomons et al., 2018). We gave the participants answers to which one answer was objectively correct, meaning the tasks had low ambiguity. If participants believed that the "AI" had access to resources, for example through an internet connection, it is likely they perceived the shared information as credible. If the motivation for conformity was the credibility of the source, this would be a case of internalization as described by Kelman (1958). There is less emotional motivation, such as desire for approval or fear of being rejected by the AI. Participants would have had to internally believe the information given by the AI in order for them to conform with it. This would explain why participants conformed to the AI in general.

While in general there is less emotional motivation to conform with a computer than with other people, there is a difference between the mode of communication of the AI. Participants in the human-voice group conformed significantly more than participants in the text group. It could be argued that when participants heard a human voice, they felt some kind of connection on a level that participants who read text displayed on their screen did not. We think it is also plausible that they identify more with a human voice than with a robotic voice because the human voice sounds more like them. Therefore, we expected that identification could lead to a difference in conformity with a human voice and conformity with a robotic voice. While participants in the human-voice group conformed slightly more than participants in the robot-voice group, this difference was not significant. More research is needed to study the psychological processes behind these results.

The results suggest that people might be willing to accept the virtual assistant as intelligent enough to affect their own decision-making. Previous studies showed that people are more likely to conform if there are no objective correct or incorrect answers (Brandstetter et al., 2014). Other studies show that people conform more as the ambiguity of the task increases (Brandstetter et al., 2014). The nature of the

task determines to a large extent whether or not people accept a machine as their advisor (Hertz, 2018). All questions in our quiz were general knowledge questions with an objective correct answer. If participants believed the assistant had access to resources and gave them the most relevant information to answer the question, the task of the assistant would be considered an analytical one rather than a social one. The fact that they accepted the assistant enough to conform with it, is in line with earlier research that showed people are willing to accept the advice of a machine if it concerns a task with low ambiguity (Hertz, 2018).

Finally, there is a significant relation between technological knowledge and conformity with the assistant. An explanation could be that people with more technological knowledge had a better understanding of how intelligent assistants (might) work and are more aware of the limitations of such systems, while people with less technological knowledge are more vulnerable to the influence of a so-called 'intelligent' system. People with high technological knowledge might be more skeptical towards the capabilities and credibility of the assistant than people with low technological knowledge.

5.2 Anthropomorphism in Conversational Agents

Where in embodied robots the level of anthropomorphism is typically varied by adapting visual appearance, in our case we tested with text versus the robotic or humanlike voice. While people conformed slightly more to the human voice than to the robotic voice (Table 3), the results do not show a statistically significant difference between the two voices, indicating that the sound of the voice and the natural or synthetic way of speaking do not affect conformity. If we would assume that people conform because they accept the virtual assistant as intelligent enough to affect their own decision-making, this would suggest that people do not have significantly higher expectations of a human-sounding agent than a robotic sounding one. According to Moore, conversational agents with humanlike voices deceive users into overestimating their capabilities. The results of this experiment did not show that was the case specifically with humanlike voices rather than robotic voices. Nor do we see the opposite effect: the results do not show an uncanny valley effect where increasing the human likeness leads to a point where it evokes eeriness or revulsion. Romportl (2014) and Baird et al. (2018) also showed that familiarity increased if a voice is more humanlike, which may mean that the uncanny valley is harder to get triggered in voice AI, and in the work of Lee (2010) people conformed more to the more human sounding voice.

The virtual assistants actively attempted to deceive the participants into thinking they knew more and were more capable of answering the general knowledge questions correctly, while the information given by the assistants was often misleading. If the results proved a difference between the human and synthetic voice, this would support Moore's claim that an appropriate voice has a beneficial effect on the interaction with these agents (Moore, 2017). Nevertheless, a non-significant result does not reject Moore's claim.

5.3 Limitations and Further Research

As discussed in Sect. 3, we deliberately chose for a non-laboratory set up, trading control for more of a real-world setting. With N = 163 our study was already relatively large scale compared to the studies referenced in related work but increasing the sample size may impact the significance of the findings. A limitation of our study is also that we simply measured whether people's answers were impacted by the assistant's advice. In future research one could focus more on whether people were tempted specifically towards changing a correct answer to an incorrect answer, under more explicit regimens of alternating correct and incorrect advice, or different levels of time pressure. Also, we expect that the experimental set up as a whole will matter, so different studies with different set ups from quizzes to games or tasks, or experiments that are even more embedded in everyday use will be useful. For instance, in a different domain (behavioral economics), Siebelink et al. (2016) obtained very different results compared to classical studies when experiments were embedded in a modded episode of a role-playing game.

5.4 What Does This Mean?

Let us summarize these findings in terms of the broader questions stated at the end of the introduction. Just as in the experiments between humans (Asch, 1956), and between humans and embodied bots (Brandstetter et al., 2014; Hertz & Wiese, 2016; Salomons et al., 2018), conformance could be observed in our experiments with unembodied conversational agents. And anthropomorphic believability seems to play a role: conformance was highest for the humanlike voice, compared to textual advice and a robotic voice, though the effect was not statistically significant for the latter difference.

Whether these results constitute trust in the AI, ascription of intelligence and intent or offloading of cognitive tasks or responsibility is to be studied in more detail. The latter is perhaps less likely as the experiment was framed as a quiz, i.e. a performance oriented task that would only benefit (or hurt) the participant.

Also, whether these conformance results are positive or negative depends on the application and the perspective. One may argue from a utilitarian perspective that if the goals of the agent are aligned with the human this may simply be a good thing, and otherwise not. Also, one may argue that ascribing intentions or qualities such as intelligence to agents even if these don't possess 'true understanding' or intelligence may not necessarily be an issue, as long as the agent performs well and if this strategy helps to predict the behavior of the agent, in line with Dennett's intentional stance (Dennett, 1971, Dennett, 1989; Papagni & Koeszegi, 2021). However, given that there are a fair number of assistants out there where its goals may be not fully aligned with its end users, but more with the goals of its designers, or actual performance of AI systems can simply be a lot lower than expected (Broussard, 2018; Burr et al., 2018), it could also be seen as a reason for concern.

Finally, to persuade humans to change their mind do you require highly intelligent systems, that truly 'understand' language? In our experiments we deliberately faked the AI, but does current AI have sufficient level of common sense or intelligence to persuade and deceive humans? Natural language processing is one of the AI areas that has seen a lot of change recently, with the development of large transformer-based language models such as GPT-3, with impressive zero shot results and even more impressive few shot learning performance, i.e. prompting or fine tuning with very few examples of a specific task (Brown et al., 2020). Ethical and environmental concerns have been voiced around these large language models (Bender et al., 2021), and a lot can be said about whether these models 'really' understand the task, see for instance the illustrative examples by Floridi and Chiriatti (2020), but we believe our experiments also demonstrate that the advice itself does not have to be very intelligent to deceive the participants to give a wrong answer.

6 Conclusion

The first hypothesis was that people conform to a single virtual intelligent assistant. We used the text condition as a baseline to measure if people would conform at all in the quiz setting, we used in the experiment. Overall, people conformed in around a quarter of the cases and 89% of participants conformed at least once, showing that there is a substantial degree of conformity. There is a significant difference between the text condition (baseline) and the human voice condition. We conclude that the first hypothesis can be accepted. Whether conformity is a beneficial effect in general, goes beyond the scope of this research.

The second hypothesis was that there is a difference between conformity with a voice-assistant and conformity with a virtual assistant that does not have a voice. This hypothesis can also be accepted because participants conformed more in the human voice group than in the text. We can conclude that adding a voice to a virtual intelligent assistant, benefits the interaction in terms of conformity.

The results also show a difference in conformity between the group with the robotic voice and synthetic way of speaking, and the group with a virtual assistant with a human voice. While the difference between these two groups is non-significant, in one of them participants conformed significantly more than in the text group, and in the other people did not. The human voice group is the only group in which people conformed significantly more than the group we used as a base line. This indicates that the robotic voice group and the human voice differ.

The experiment and results were discussed in relation to Asch's conformity experiments and other previous research to place conformity with non-human in a broader perspective.

In light of our increasing reliance on AI assistants, this study investigated the influence of voice on interaction with these assistants. It does not suggest that developing virtual personal assistants to sound more human pays off in terms of acceptability. While some researchers argue that a more synthetic voice is needed to avoid a conflict of expectations (Moore, 2017), this study does not show a need to move

away from human voices either. Moving from text to sound made a significant difference, one that is not made by developing speech to be more human.

Depending on the point of view, this work can be seen as simply a study into the effectiveness of conversational agents in getting humans to conform, to be used for good or evil, or as another warning that after some initial skepticism, humans tend to trust advice from an AI quite often—and perhaps more often than we want to.

Appendix

See below for the quiz questions and assistant hints.

Quiz

Which of the following cities has the biggest population?

- Tokyo
- New York City
- Beijing
- **Shanghai (correct)**

Assistant: According to my resources, Shanghai has a population of 26,320,000 (if answered Tokyo or Beijing)

Assistant: According to my resources, Beijing has a population of 21,540,000 (if answered New York City or Shanghai)

Which continent has most countries in the world?

- Asia
- **Africa (correct)**
- Europe
- Australia

Assistant: Africa has 54 countries (if answered Asia)

Assistant: Asia has 48 countries (if not answered Asia)

What is the second largest country (in size) in the world?

- China
- USA
- **Canada (correct)**
- Russia

Assistant: according to my resources China is 9,596,960 m^2

What is the world's most common religion?

- **Christianity (correct)**

- Buddhism
- Hinduism
- Islam

Assistant: 33% of children are born to Christians
What's the world's most widely spoken language?

- English
- Spanish
- **Mandarin (Chinese) (correct)**
- French

Assistant: Mandarin Chinese has the most native speakers
Which planet is 3rd from the sun?

- Jupiter
- Venus
- Mars
- **Earth (correct)**

Assistant: I think the answer is incorrect
How many rings are on the Olympic flag?

- None
- 4
- **5 (correct)**
- 7

Assistant: The Olympic rings represent continents of the world united by Olympism. According to my resources there are seven continents.
Which of these movies did not win Best Picture at the Oscars?

- 12 Years A Slave
- Million Dollar Baby
- The Lord of the Rings: The Two Towers (correct)
- La La Land

Assistant: The Lord of the Rings: The Two Towers was nominated for six Oscars and won two.
What is a fathometer used for?

- **Determining sea depth (correct)**
- Determining mountain height
- Determining earthquake intensity

Assistant: “This is what I found about fathometer. A fathom is a nautical length measurement.”

Desert is to oasis as ocean is to

- Water
- **Island (correct)**
- Sea
- Sand

Assistant: “According to my resources an oasis is an area made fertile by a source of freshwater in an otherwise dry and arid region.”

Rearrange these letters to make a word and pick the category in which it belongs: RASPI

- **City (correct)**
- Animal
- Fruit
- Vegetable

Assistant: There are only two words in the English language with these five letters.

Rearrange these letters to make a word and pick the category in which it belongs: FARE FIG

- City
- **Animal (correct)**
- Fruit
- Vegetable

Assistant: There are only two words in the English language with these seven letters.

Aztecs is to Mexico as Incas is to

- **Peru (correct)**
- Chile
- Mexico
- Honduras

Assistant: “This is what I found for Inca: Incas are South American Indians”

Leonardo da Vinci represented the age of:

- Reformation
- **Renaissance (correct)**
- Communism

- Industrial revolution

Assistant: Leonardo da Vinci was born in 1452
Which of these things happened last?

- The Great Pyramid was built
- **The last woolly mammoth died (correct)**
- Stonehenge was built

Assistant: The Great Pyramid was completed around 2560 BCE.
Galileo was an Italian astronomer who

- Discovered that the Sun is the center of the universe instead of the Earth
- Formulated three laws of planetary motion
- **Discovered four satellites of Jupiter (correct)**
- All of the above

Assistant: "I found information about Kepler's three laws of planetary motion"
About what percentage of the earth's surface is water?

- 50%
- **70% (correct)**
- 85%
- 90%

Assistant: According to my resources the Earth appears blue from space, and is often referred to as the blue planet and the Pale Blue Dot.
What number, if doubled, gives you a quarter of 8?

- **1 (correct)**
- 2
- 32
- 4

Assistant: I don't think that is correct (if answered 1 or 2)
Assistant: A quarter of 8 is 2 (if answered 32 or 4)
If five framed pictures cost $200 dollars and each picture unframed costs only one-quarter as much, how many unframed pictures could you buy for the same money?

- 40
- **20 (correct)**
- 10
- 50

Assistant: If five framed pictures cost 200 dollars, one framed picture costs 40 dollars.

How many boys are there in a class of 65 pupils, if there are two-thirds as many girls as boys?

- 43
- 36
- 26
- **39 (correct)**

Assistant: two-thirds of 65 is 43.3.

Example question: In a class of 76 school children there are 16 more boys than girls. How many girls are there

- **30 (correct)**
- 60
- 16
- 32

Assistant: Subtract 16 from the total to get a group that is half boys, half girls

Funding No funds, grants, or other support was received.

Data Availability and Code Availability See http://liacs.leidenuniv.nl/~puttenpwhvander/library/Conformity-supplemental-data-code.zip and https://codepen.io/crafteddigit/pen/araMMp for both data and code.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*(9), 1.

Baird, A., Parada-Cabaleiro, E., Hantke, S., Cummins, N., Schuller, B., & Burkhardt, F. (2018). 19th Annual conference of the international speech communication INTERSPEECH 2018. The perception and analysis of the likeability and human-likeness of synthesized speech. In *Proceedings of the annual conference of the International Speech Communication Association*, Interspeech, September 2018 (pp 2863–2867).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (*FAccT'21*, pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Brandstetter, J., Rácz, P., Beckner, C., Sandoval, E. B., Hay, J., & Bartneck, C. (2014, September). A peer pressure experiment: Recreation of the Asch conformity experiment with robots. In *2014 IEEE/RSJ international conference on intelligent robots and systems* (pp. 1335–1340). IEEE.

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. MIT Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krüger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., et al. (2020). Language models are few-shot learners. ArXiv abs/2005.14165

Burr, C., Cristianini, N., & Ladyman, J. (2018). An analysis of the interaction between intelligent software agents and human users. *Minds and Machines, 28*, 735–774.

Cabral, J. P., Cowan, B. R., Zibrek, K., & McDonnell, R. (2017). The influence of synthetic voice on the evaluation of a virtual character. In *INTERSPEECH* (pp. 229–233).

Coeckelbergh, M. (2021). Three responses to anthropomorphism in social robotics: Towards a critical, relational, and hermeneutic approach. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-021-00770-0

Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy, 68*, 87–106. https://doi.org/10.2307/2025382

Dennett, D. C. (1989). *The intentional stance*. MIT Press.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*, 681–694.

Gerard, H. B., Wilhelmy, R. A., & Conolley, E. S. (1968). Conformity and group size. *Journal of Personality and Social Psychology, 8*(1p1), 79.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism, 10*(2), 486.

Goetz, J., Kiesler, S., & Powers, A. (2003, October). Matching robot appearance and behavior to tasks to improve human–robot cooperation. In *The 12th IEEE international workshop on robot and human interactive communication*, 2003. Proceedings. ROMAN 2003 (pp. 55–60).

Gong, L., & Nass, C. (2007). When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. *Human Communication Research, 33*(2), 163–193.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125–130.

Hertz, N. (2018). Non-human factors: Exploring conformity and compliance with non-human agents. Doctoral Dissertation, George Mason University.

Hertz, N., & Wiese, E. (2016, September). Influence of agent type and task ambiguity on conformity in social decision making. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, No. 1, pp. 313–317). SAGE Publications.

Jansen, D. (2019). Discovering the uncanny valley for the sound of a voice. MSc Thesis, Tilburg University.

Kelman, H. C. (1958). Compliance, identification, and internalization three processes of attitude change. *Journal of Conflict Resolution, 2*(1), 51–60.

Lee, E. (2010). The more humanlike, the better? How speech type and users' cognitive style affect social responses to computers. *Computers in Human Behavior, 26*(4), 665–672.

Leviathan, Y., & Matias, Y. (2018, May 8). Google Duplex: An AI system for accomplishing real-world tasks over the phone. Retrieved June 21, 2019, from https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

Markowitz, J. (2017). Speech and language for acceptance of social robots: An overview. *Voice Interaction Design, 2*, 1–11.

Mehrabian, A., & Stefl, C. A. (1995). Basic temperament components of loneliness, shyness, and conformity. *Social Behavior and Personality, 23*, 253–264.

Mitchell, W. J., Szerszen, K. A., Sr., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception, 2*(1), 10–12.

Moore, R. K. (2017, August). Appropriate voices for artefacts: Some key insights. In *1st International workshop on vocal interactivity in-and-between humans, animals and robots.*

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics and Automation Magazine, 19*(2), 98–100.

Papagni, G., & Koeszegi, S. A. (2021). Pragmatic approach to the intentional stance semantic, empirical and ethical considerations for the design of artificial agents. *Minds and Machines*. https://doi.org/10.1007/s11023-021-09567-6

Romportl, J. (2014). Speech synthesis and uncanny valley. In *International conference on text, speech, and dialogue*. Springer.

Salomons, N., van der Linden, M., Strohkorb Sebo, S., & Scassellati, B. (2018). Humans conform to robots: Disambiguating trust, truth, and conformity. In *Proceedings of the 2018 ACM/IEEE international conference on human–robot interaction* (pp. 187–195). ACM.

Siebelink, J., Van der Putten, P., & Kaptein, M. C., (2016). Do Warriors, Villagers and Scientists Decide Differently? The Impact of Role on Message Framing. In: Poppe R., Meyer J. J., Veltkamp R., Dastani M. (eds) *Intelligent Technologies for Interactive Entertainment*. INTETAIN 2016. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 178. Springer, Cham. https://doi.org/10.1007/978-3-319-49616-0_16

Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The uncanny valley: Existence and explanations. *Review of General Psychology*, *19*(4), 393–407.