Differences and similarities in perceptions of interactions with Artificial Social Agents between German and English speakers

Boleslav Alexandrovic Khodakov[1]

Supervisor(s): Willem-Paul Brinkman[1]**, Nele Albers**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Boleslav Alexandrovic Khodakov
Final project course: CSE3000 Research Project
Thesis committee: Willem-Paul Brinkman, Nele Albers, Odette Scharenborg

Abstract

Humans interact with various Artificial Social Agents (ASAs) on a daily basis. ASAs range from the Honda robot ASIMO to Apple's Siri. To measure the perception of human-ASA interactions, a standardized questionnaire was created. Yet, this questionnaire was so far only available in English and Chinese. It has been found that culture can affect how these interactions are perceived. The aim of this study is to answer the question: **What are the differences and similarities of the English and German human-ASA interaction interpretations?** In this paper, we translate the questionnaire into German, validate it. Once proven valid, we give the English and German questionnaire on bilingual participants who watch a human-ASA interaction video and rate it in both languages. We measure the differences and similarities between the English and German responses. At the end, we combine the finding from the questionnaire results with examples from literature to form recommendations for future ASA developments. We conclude that an average good level of correlation between the two languages for the 90 questionnaire items (ICC $M = 0.65$, $SD = 0.14$, range [0.27, 0.90]), on the construct level (ICC $M = 0.8$, $SD = 0.1$, range [0.51, 0.92]), and for the 24 representative items ($M = 0.67$, $SD = 0.14$, range[0.31, 0.90]). Additionally, we found systematic differences between the English questionnaire scores of the bilingual sample seen in this study and a previously established mixed-English sample.

1 Introduction

Artificial Social Agents (ASA) are a widely used technology in the modern day world – from simple chatbots to physical robots, ASAs can be found in a plethora of environments [8]. ASAs can be used for a wide range of scenarios, for instance as tour guides [21] or a medical assistants used to help people quit smoking [23] [1].

There have been several studies examining how what makes the ASA social [4], what makes it an effective assistant [23] [1]. Even the preferred degree of extroversion of an ASA has been studied [33], leading to the conclusion that an in-between solution is best.

However, culture may affect people's perception of ASAs [30] [31] [28]. Specifically, take the example of German culture compared to English culture. It has been shown that among other things, German communication styles differ from English ones [14] [13]. For example, the style of politeness varies significantly in requests made by German and English speakers [29]. Even when dialects of (Austrian) German are evaluated, clear differences are observed between different varieties [21]. Several attempts have to localize AI can be observed, from the CANVAS framework [32] to even sub-national regulation policies in case of Germany [26].

It becomes clear that the issue of ASAs is an extensive one - one which will only become larger as the field of Artificial Intelligence gets more integrated into people's daily lives. To account for future progress of the technology researchers and ASA producers need to be able to objectively measure human-ASA interactions. In Fitrianie et al. [8] the creation and validation of a universal human-ASA questionnaire can be observed. The questionnaire is available in English, a universal language. Yet, for non-native speakers it is an additional layer of interpretation. Thus, in Li et al. [24] researchers attempted to translate the questionnaire into Mandarin.

We will take the previous efforts one step further. The work presented is intended to answer to question: **What are the differences and similarities of the English and German human-ASA interaction interpretations?** In this paper we will outline our efforts of translating the English human-ASA questionnaire into the German language. Upon creation of the translation, we evaluated its correlation to the original and improved it to the degree where it can be considered a German equivalent of the original. Afterwards, we conducted a large *(n $\geq$ 72)* survey based on this questionnaire to gather data from bilingual German-English speakers (with German as the primary language). We evaluate the differences and similarities between their answers of the English and German versions of the questionnaire. We also compare the English subset of data to already existing data from mixed-English respondents. Lastly, we discuss recommendations for future ASA development based on a combination of data obtained and a literature study.

2 Background

This section is concerned with providing the reader with background on two aspects of the study: Human-ASA interaction.

Unfortunately, little literature exists on the precise topic of this paper.

2.1 The ASA questionnaire

The motivation behind this paper comes from the work of Fitrianie et al. [8]. The research community behind the study developed a standardized questionnaire, known as the "ASA questionniare". The questionnaire consists of 90 items, which is a results of refining previous studies (i.e. Fitrianie et al. [11] Fitrianie et al. [10]). The items were proven to have a good level of convergent and discriminant validity.
In Fitrianie et al. [10] the term ASAs was clearly defined as "computer-controlled entities that can autonomously interact with humans following the social rules of human-human interaction". The same definition applies to the current study.
Later, in Fitrianie et al. [11] a world model of human-ASA interaction was created. The focus of the model was human-ASA interaction. Thus, some variables such as users' previous relations with ASAs and their environment were not controlled. The 189 possible constructs resulting from previous

studies, 19 were chosen to represent roughly 80% of the total constructs.
In Fitrianie et al. [9], a generated set of validated questionnaire items for these constructs was established to be reliable. The reliability was measured by 192 participants using these items to measure a human-ASA interaction of people with the Honda robot ASIMO. After analyzing factor analysis models, the researchers found the remaining 90 questionnaire items had a good level Fitrianie et al. [8].

2.2 The Chinese version of the ASA questionnaire

In Li et al. [24], an effort was taken to create and validate a Chinese version of the original ASA questionnaire (similarly to this study). The study bases itself on the work of subsection 2.1. It also provided us with an overview of the original studies.
The Chinese translation study consisted of three formative bilingual assessments, and one summative one. Firstly, 101 Chinese translations of English items (created by bilingual researchers) were evaluated by bilingual participants in the first formative round. In the next formative cycle translations were evaluated again. 53 were new, as the previous versions had low correlations scores. For the third formative round, only 39 items with remaining low correlation scores were re-evaluated.
Three new experts additionally back-translated the Chinese questionnaire into English without having the original English items. Comparing the two English versions, 5 items were identified as having discrepancies, and new translations were formulated. This resulted in the final translation.
Finally, in the summative round 242 participants evaluated one of 14 human-ASA interactions (shown through video) in both English and Chinese. It was found that the that the intraclass correlation coefficient (ICC) values had a good level on both item- and construct-level. The researchers found a postivie bias from the Chinese questionnaire for four constructs, negative bias for two. On item-level, eight positive and three negative biases were present.
The study also compared their results to a previous mixed-English result. They found that the Chinese-speaking participants gave higher scores overall for the constructs: Agent's Appearance Suitability (AAS), Performance (PF), User Acceptance of the Agent (UAA), Agent's Enjoyability (AE), User's Engagement (UE), and Attitude (AT).

This study's approach is inspired by Li et al. [24] research.

3 Questionnaire translation

The following section discusses the process of translating the original human-ASA interaction questionnaire from Fitrianie et al. [8] into the German language. Figure 1 illustrates the several steps described in the next subsections.

3.1 Step 0: Receiving approval

Ethics is an important part of conducting studies, especially ones involving human participants. Before we started, our supervisor requested approval for the study from TU Delft's ethics committee by filling in Human Research Ethics Checklist (HREC) form. Approval was granted. Additionally, an

Figure 1: Visualization of questionnaire translation, validation efforts. The first three bubbles from the top are about translation.

Open Science Framework (OSF) form [18] was filled in before the steps beneath were conducted. It ensured that we thoroughly thought about the procedures of the research before conducting it, and documented it.

3.2 Step 1: Translation

Experts from RWTH Aachen, fluent in both German and English and experienced with artificial agents, translated the original ASA questionnaire into the German language. One of the experts combined the translations into a single file. In case of multiple translations for some English items, all versions were provided. The translations were sent to us.

3.3 Step 2: First cycle of formative bilingual assessment

Along with my colleague, we divided the questionnaire into two halves, one for each person to work on. The split divided the items as evenly as possible. The first half consisted of the first 12 constructs of the questionnaire, the second half of the remaining 12. This step was necessary to reduce participants' fatigue. 44 English items (items HLA1 - AE4) with their corresponding 50 German translations were part of the first half. The second half consisted of 46 English items (items UE1 - UAI4) and their 50 German translations. The division is the same as in Li et al. [24].

We recruited bilingual participants *(n=30, per each half)* from the online crowd-sourcing platform Prolific Academic. Participants rated an interaction of the Honda robot ASIMO

with two TV-show hosts in a 30-second video clip. The survey consisted of both English and translated German items from either the first or second half of the human-ASA questionnaire. Each participant answered both German and English questions.

Before the start of the main Qualtrics survey, participants were required to answer several consent questions on the platform. Participants only had access to the survey, if they consented to each item in the consent form.

Lastly, to assure each participant was answering the questionnaire truthfully (instead of just clicking randomly), 14 attention checks (7 in English and 7 in German) were added to each survey. If a participant failed any attention checks, their results were not considered for the study (they were not counted towards n).

3.4 Step 3: Translation evaluation, repetition

The results from subsection 3.3 were evaluated. To evaluate the results the programming language R was used. We modified the codebase from Li et al. [24] as it was used for the same calculations. The main package used was `nlme`. The anonymized survey results from Qualtrics were exported into `.sav` files. These files were first transformed from raw data into the essential data. In case there were multiple German translations for English items, the respectable English items were duplicated. This meant that for a 1:1 relation - each German translation was tied to its own English item.

After data transformation, intraclass correlation coefficient (ICC) values for individual items were calculated, similarly to Li et al. [24]. ICC values show the correlation between the English items and their German counterparts. Cicchetti [3] provides guidelines for interpretation of the ICC values. ICC values less than 0.4 are considered poor, between 0.64 and 0.74 values are labeled as good. Anything above 0.75 (upto 1.0) is considered excellent. An alternative suggestion by Koo et al. [20] is: below 0.5 is poor, between 0.5 and 0.75 is moderately reliable, between 0.75 and 0.9 is good, above 0.9 is excellent. Lastly, Mehta et al. [27] reports value of at least 0.6 being required to represent substantial reliability in scale studies. Thus, it was chosen to take ICC values of 0.6 as the cut-off point for per-item level evaluation.

Any translated items which had poor correlation with their English counterparts were dropped. For items with multiple translations, it was sufficient for one translation to be "good". At the end of the survey, participants were asked whether they would recommended the use of their data for (this) research. 3 participants did not recommend using their data. As a result, we also calculated ICC values for recommended data only (57 participants). Translations whose ICC scores were previously higher than 0.6, but dropped after using only recommended data, were also dropped.
The total amount of items which had to be re-translated (for both halves) was 35. 2 out of 35 items were only low-ICC when using recommended data exclusively. The items were once more sent to the German experts for a revised translation. Then for the items (and their translations), steps from subsections 3.2-3.4 were repeated.

3.5 Step 4: Second cycle of formative bilingual assessment

The current step is repeating sections 3.3, 3.4 for the new translations of the 35 items whose translations had low ICC scores in the first survey.

The process was largely the same. Another Qualtrics/Prolific survey was set up for the 35 items and their 78 translations. Each item had at least 2 translations, to increase chances of finding a high-ICC one. The survey was not divided based on constructs, all 113 items were in one survey. Its length was similar to surveys from round 1. 30 bilingual participants rated ASIMO's interaction with humans (the same video). 29 participants recommended using their data for research.

Having analyzed the new data, 11 items were found to have no high-ICC translations. One of those only had a low-ICC score when using recommended data exclusively. Additionally, it was found that due to a rounding error, 1 previously accepted item from round 1 also had to be re-translated. Thus, 12 items were sent to the German experts for a re-translation.

With the new translations for round 3, the German experts also sent translations for a thirteenth item. They found its previous translation to be "too similar to another translation". It was included in round 3. However, if no translation would have a higher ICC score than the previously accepted one, the original "similar" translation would be kept.

3.6 Step 5: Final cycle of formative bilingual assessment

The current step is once more repeating sections 3.3, 3.4, 3.5 for the new translations of the 12 items whose translations had low ICC scores in the first survey. An additional thirteenth item is also added for re-translation (cf. section 3.5).

Again, the same process was applied. 30 participants, recruited through Prolific, answered a survey on Qualtrics. The survey was not divided. Due to the deficit of English items (13) compared to German ones (52), English attention checks were reduced to 3.

After evaluation it was found that 7 of 13 items had at least one high-ICC translation. Given time constrains, we chose the pragmatic approach for items without high-ICC translations. The best possible (highest-ICC value) translation was chosen, and the items were included in the summative survey.

Despite minor shortcomings, it was determined that 3 translation rounds were the maximum allowed due to time constraints related to the project. We are satisfied with having a $\frac{84}{90}$ translation success-rate.

3.7 Different gendered forms in German

We were made aware by the supervisor that gender is currently a highly debated topic in Germany. Unlike English, German is a gendered language. It was agreed together that for highest inclusivity, multiple forms of translations will be created. Thus, the same general translations have the following forms: **male, female and plural** forms for humans, and **male, female, neutral** *(i.e. "er/sie")* forms for ASAs. These versions were used with according videos in the summative survey. For the formative survey, the specific case of multiple humans (plural form) and male robot was used.

4 Methods

This section is about finding differences and similarities of English and German speakers in terms of their rating of human-ASA interactions through the conduction of a summative survey.

4.1 Design and procedure

Before conducting the study, ethical approval was granted for this research by TU Delft. Additionally, a separate OSF form [17] was submitted, assuring the research was conducted responsibly (cf. section 3.1).

For the summative assessment, bilingual German-English participants (n = *72-82*), with German as their primary language, were recruited to rate one of 14 videos from different human-ASA interactions.

Each participant was randomly assigned one of 14 available videos (further outlined in subsection 4.3). The agents are exactly the same as in Li et al. [24], but the videos sometimes differ. Each clip had a duration of roughly 30 seconds.

To control for fatigue effects, the same principal as in section 3.3 was used: 90 items and their German counterparts (total 180 items) were split into two sub-questionnaires of 88 and 92 items each. 14 attention checks were added into each sub-questionnaire, 7 in English and 7 in German. This ensured only truthful responses were recorded, as only the surveys of participants who answered all checks correctly were recorded. Participants rated English and German items of either the first, or second half.

Before participants were allowed to fill in a sub-questionnaire, they had to fill in a consent form. Their internet browser was also checked for compatibility - they had to watch a video and answer a control question about its content.
Then, the actual ASA-interaction video played, and participants were allowed to begin submitting responses. The ASA-interaction video was re-watchable at all times during the questionnaire. Only upon answering all questions could participants submit their response.

4.2 Participants

The authors from Li et al. [24] have found out in their study that a sample size of 110 participants was for detecting a small effect (d = .2) with an 80% chance.

This study goes by the same logic. Having two groups (due to two questionnaires), requires a doubling of the size. Adding a small safety margin it was chosen to have 120 total participants for the questionnaire. Similarly to section 3.3 the participants were once more recruited via Prolific Academic. The questionnaire being hosted on the platform Qualtrics. Participants were paid the minimum amount allowed by Prolific *(6£ per hour)*. We have recorded participants self-reported age, self-reported gender, self-reported highest education. Before the data was handled, it was anonymized.

Unfortunately, due to a late launch of the questionnaire and time constraints, only 72 participants answered the first-half survey, with 82 answering the second-half one. The analysis code requires an equal amount of participants per each half. Thus, 10 participants' entries were removed arbitrarily from the second-half survey, resulting in 72 participants per survey. The results used in the study are based on the first data collected between the 19th and 22nd of June. We are still aiming for 120 participants per questionnaire part for future work.

4.3 Materials

Adhering to Li et al. [24], 14 videos used in that study were also used for the purposes of this research. The videos cover a wide amount of situations, and should account for all constructs and dimensions of the ASA questionnaire (see Li et al. [24]). The videos had the humans interact with the following ASAs: iCat, DeepBlue, Amy, Furby, Siri, HAL 9000, Poppy, Sim Sensei, CHAPPiE, Aibo, Sarah, Nao, Marcus, and a dog.

In addition, similarly to Li et al. [24], the study limited the ASA questionnaire to the third person version (a first person version also exists in English). This simplifies the task of the participants. It also distances the participants from the role of the people interacting with ASAs in the video.

4.4 Data preparation and analysis

Given the similarity of this study to Li et al. [24], the following section resembles section *METHODS, Data preparation* from the aforementioned paper.

To analyze the data collected the statistical programming language R (v.4.2.3) was used. This study's R scripts are modified versions of R scripts from Li et al. [24], for they serve the same purpose. The study followed the approach from Finch et al. [7] for calculating ICC scores. For each of the 24 constructs a multilevel model was fitted on its items with fixed intercept and participants used as random intercept. The same process was applied to the individual 90 items. The similarity of participants' English and German ratings were compared. In scientific terms, the proportion of total variability in score ratings that was attributable to an individual participant was examined. The R package `nlme` (v.3.1-162) was used for this process. For calculating the scores of the 24 constructs, the mean of each the respectable constructs' items for each participant were taken. During the analysis, the 24 representative items of the short ASA

questionnaire version were also analyzed.

To estimate the mean, standard deviation, 95% Credibility Interval (CI) of the posterior t-distribution of the mean differences in the score of both languages, the R package `BayesianFirstAid` (v.0.1) with its *t*-test capabilities was used. 95% CIs which did not include zero were regarded as credible indication of a systematic positive/negative bias and requiring future conversion correction in the future. Broad priors were used in the analysis, as outlined in Kruschke [22]. For credibility intervals, a 95% highest posterior density interval was used. This is the narrowest interval containing 95% of the probability mass.

Lastly, systematic differences in English questionnaire scores between the bilingual sample and a previously collected mixed-international English-speaking sample from Fitrianie et al. [8] were investigated. The authors of Fitrianie et al. [8] collected data from the same 14 agents used in the current study (different videos), and even used the same platform to find participants (Prolific Academic). The sample might include German-speaking participants, though this is not explicitly reported. The sample from Fitrianie et al. [8] was thus regarded as simply mixed-English - the differences between those results and ones collected in this study are attributed to culture.
The R package `Rethinking` (v.2.31) was used to follow a Bayesian approach. A multilevel model with a Gaussian distribution was fitted on each construct score, with a linear model that included culture as a fixed effect and agent as a varying effect with partial pooling. Uninformed priors were used in the analysis. For the interpretation, we regarded 95% CI of the culture coefficient estimate that excluded zero as a credible indication of a difference between the two sample groups. Posterior probabilities of either positive or negative bias between two sample groups were also calculated. This was done by taking the posterior distribution area that was either small or greater than zero, whichever was larger.

All data sets, analysis scripts, and outcomes files are online available (see section 8.6).

5 Results

Given the similarity of this study to Li et al. [24], the following section stylistically resembles section *RESULTS* from the aforementioned paper.

5.1 Correlation between English and German ASA Questionnaire

A good correlation level was shown by the mean ICC value of the 24 constructs and related dimensions (ICC $M = 0.8$, $SD = 0.1$, range [0.51, 0.92]), as well as the 90 questionnaire items (ICC $M = 0.65$, $SD = 0.14$, range [0.27, 0.90]). Presented in table 1 are the results: On an item-level 64% had a good or excellent correlation, while on a constructs level this was the case for 91% (Table 2). Additionally, with an average ICC value of 0.67 ($SD = 0.14$, range[0.31, 0.90]), a good correlation level was found for the 24 representative items of the short version of the ASA questionnaire(Table 3). For 18 of these items (75%), the correlation level was good or excellent.

5.2 Variation between English and German ASA questionnaire

The mean score differences between the English and German questionnaires are estimates for score equivalence between the two languages, and for positive biases (i.e. the German score being higher than the English one) or the negative biases (the inverse). For the 24 constructs, Table 2 shows a grand mean difference in absolute terms of 0.11 and a grand mean of standard deviation (SD) of 0.089. Scores are in the range of [-0.14; 0.52]. For three constructs, the credible interval was above zero. Thus, there was a positive bias. No negative bias was found, as there exists no credible interval which was below zero.
Similarly, the item level can be analyzed. Table 3 shows this for the 24 representative items. A grand mean of 0.11 can be seen ($SD = 0.09$). Scores are in the range of [-0.08, 1.13]. One item shows credible indication of positive bias.
For the complete set of 90 items (Grand Absolute Mean 0.11, SD = 0.09, range [-0.44, 1.13]), 7 items have positive bias, while 4 items show negative bias (Table 4).

5.3 Comparison of human-ASA interaction between different cultural backgrounds

Table 5 depicts the results of the construct score analysis between the German primary-tongue sample and the mixed-international English-speaking sample. We found three credible indications of a difference. In all these cases, the posterior probability was above 99%. Across 14 ASAs, the German primary tongue sample gave a higher score for the constructs APP (Agent's Personality Presence), SP (Social Presence). The same sample gave a lower score to construct IIS (Interaction Impact on Self-Image).

6 Discussion

In this section results are compared to those of previous work and placed in a broader context. More specifically, we will form recommendations for future ASA development based on the combination of data obtained in the practical part of this paper, as well as a literature study.

6.1 ASAs in functional context

Before providing future recommendations for ASAs, it is important to state how ASAs can generally be used. The information provided in this subsection aims to tell the reader about some functional insights of ASAs.

Firstly, Li et al. [25] studied distinct ways of humans collaborating with AI based on two aspects - exploitation vs. exploration, and interdependence vs. independence. The study of 1367 participants found that humans highly prefer interdependent exploitation cooperation with AI, in the context of the video game DOTA. In essence, it was found humans like to work with AI, not in parallel.

The findings are confirmed by Zamora [34], where American and Indian participants evaluated how useful chatbots were to them. The study found that less than 50% of participants would use chatbots again. This was due to the bots being slow, and of low usefulness and feeling like a middleman. From those responses, and the 22 positive ones, the study found that robots should be high performing, smart and seamless, personable and context-aware. Robots were found to be useful mostly for administrative tasks.
These general findings are important for our recommendation about future developments, since we can tighten our vision towards complementary AI.

In our future recommendations we should not only account for the style of interactions, but also the amount of information presented. Becks et al. [2] is a good basis for this. The suggested WoOZ experiment is likely to conclude that AI nudging the user is the best option. It is thought in the paper that presenting all information that AI uses will overwhelm the user. Meanwhile, if AI simply gave a solution, humans would become dependent on it.
Thus, we can suspect that we should provide complementary aid to the user in both style and amount of information.

The last important aspect of functionality is the embodiment future ASAs should adapt. Which entities would humans be more willing to accept? Hoffmann et al. [12] found that this highly depends on the context. Though the study was in an early stage, researchers still found that a physical robot (compared to a virtual one) is better perceived at tactile interaction, mobility and corporeality, while no differences were found for (nonverbal) expressiveness (according to a custom scale for the study).
We are taught to take the work environment into account for future recommendations.

6.2 ASAs in social context

On the opposite side to functionality, we have the less rational social aspect of ASAs. Though ASAs effectively simulate natural behavior, the way in which they communicate with the user is still crucial to a good human-ASA interaction. We would not want an ASA meant for children to sound rude, for instance.

The social qualities of ASAs were studied by de Graaf et. al [4]. In the study, 21 people were given a small zoomorphic Amazon-Alexa like ASA. The ASA was not very capable, being able to only answer preprogrammed questions. Yet, the point of the study was: what made the robot social? The conclusion was that an ASA should have two-way interaction capabilities, show its own thoughts and feelings, be socially aware, autonomous. It should also be cozy, show similarity to its owner and have respect for them.
Noting down these qualities will help us avoid misinterpretation of future ASAs, as communication becomes more clear.

Another case studied the exact degree of extro- or introversion that made users the most comfortable with ASAs. The study by Völkel et al. [33] evaluated which type of chatbot (extroverted, average, introverted) users preferred. On paper extroversion was preferred, yet users interacted more with an introverted chatbot. This could be due to the more intelligent phrasing used by the introverted chatbot, whilst the extroverted model was more emotionally oriented.
For us, this means ASAs should show competence through use of professional terminology without losing the "human touch" and sounding robotic.

Notably, as found by Kim et al. [19], people tend to prefer functional AI over social AI (in terms of usefulness). The social aspects of ASAs seem to have been looked less into by previous research. Yet, as ASAs develop further, these should not be overlooked.

6.3 Linguistic differences of German and English

With the functional and social aspects clarified, it is time to examine the differences between the German-speaking and English-speaking users of ASAs. Before connecting recommendations with the cultural aspect, we should examine the corresponding languages first. Language is one of the first layers of perception of any interaction, including human-ASA. For future recommendations, linguistic adaptation of ASAs towards the region of use is crucial for success.

House [14] specifically studies the differences between German and English communicative styles. This is essential to know, if one wants to adapt ASAs to the German-speaking markets. The study found several differences, such as English speakers having more conversational routines while German speakers were more context-driven. German speakers are less likely to use small talk, they are more content-oriented. There is more direct expressions in German compared to English, German speakers are more verbose. The following differences can be perceived as stark, and led some non-German participants of the study perceive German speakers as rude, even when they might not have been.

The most frequent interaction with ASAs is likely to be requests (e.g. "Alexa, ..."). Thus, their "correctness" is perhaps the single most important detail in human-ASA interaction. In Ogiermann [29] the specific differences in requests between speakers of German, English, as well as Polish and Russian were evaluated. Differing levels of directness were once more mentioned, among different types to "downgrade" (downplay) requests. German and English speakers both prefer interrogative constructions. However, German speakers downgrade with downtoners (e.g. "mal eben"), while the English kind uses consultative devices (e.g. "do you think...").

Lastly, in Krenn et al. [21] it was found that even dialects of the same (Austrian) German language are perceived differently within the same context. An ASA gave guided tour videos of Baroque State Hall in standard Austrian German, Colloquial Viennese, Dialectal Viennese. Austrian Standard was perceived as most educated, professional, also most preferred for the job. Dialectal Viennese was perceived as natural, emotional, highest sense of humor. The different voices were found to trigger different assumptions. Though, the study was not using all voices of the same gender, which

might have skewed results. Additionally, the study did not account for participants social proximity to any particular language variety.

We can conclude from our research, that German-oriented ASAs should be more concise and task-oriented. They should account for German downtoners, and perhaps even know some colloquial expressions. For the English market, we believe most ASAs have been adapted over the years.

6.4 Differences in ASA perception between German and English speakers

At last, we examine the subjective realm of culture. Culture is hard to measure, yet anyone who ever traveled knows how much it can change one's perception. This also holds true for human-ASA interactions (hence this whole paper). While measuring cultural differences was mostly done in the practical part of this paper, some literature background on the topic is still beneficial for future ASA development recommendations.

There have been multiple studies comparing people from different cultural or linguistic groups based on their interaction with ASAs.

At least two studies can be named comparing Arabic speakers with others. In Obaid et al. [28] Arabic speakers are compared with English speakers. Both got to interact with an ASA. The ASA either looked Western and spoke English or looked Arabic and spoke Arabic. There were also two settings, casual and professional. The study found that neither conveyed status nor ethnicity influenced how perceived the ASA. However, participants' cultural background did influence their subjective perception of the ASA. The stronger the perception of the agent as being a member of a participant's cultural group, the better the Agent was rated.

In Salem et al. [31] another study made a robot either speak English or Arabic. The robot also adjusted to reflect levels of politeness, directness and approach to speech of the culture it was trying to mimic. People from Western and Arab cultures were given two conversations with the robot - casual and goal-oriented. The cultural background was seen as affecting the perception of the robot. Arab participants felt like robot was more polite and competent. They were more forgiving of mistakes. Arab participants were also more likely to see the robot as an in-group member.

Having confirmed culture affects perception, we move to German examples. German students were found to judge robots from different cultures differently. In Eyssel et al. [6], two groups of students were presented with the same robot. One group was told that the robot is from Turkey, and a Turkish name was given. The other group was told the robot with a German name was from Germany. The German robot was rated more favorably and was more anthromorphized.

Thus, we can conclude that ASAs should adapt not only linguistically, but also anthropomorphically, and perhaps even inner-logically (= how they compute, how they express things).

A notable exclusion to the rule is emotional adaptation, which is not needed. Qu et al. [30] found that even culturally separate German and Chinese speakers evaluated a Chinese speaking virtual ASA's valence very similarly. The ASA held a dialogue with the participants.

6.5 Connecting analysis data with literature

Primary German speakers rating constructs APP, SP higher indicates a higher acceptance of the ASA they were presented. This might be a result of the ASAs' actions resembling social presence more appropriate in a German-speaking context. Since ASAs are best serving as administrators, and German-speakers prefer context-driven/routine-driven interactions, they may evaluate these as more social. The short time frame of the interaction may have also helped, as German-speakers like to keep interactions to the point.

The lower German-speaking rating of the IIS construct cannot be tied with related literature. We suspect this means German speakers prefer ASAs while focusing on themselves, instead of using the ASA in a group context.

6.6 Recommendations for future ASA development

Based on the literature study, and the data analysis conducted, we can provide some recommendations for the development of future ASAs, within a German linguistic context.

Linguistically, a good ASA should be context-driven and routine-based for German speakers. It should avoid small talk. The ASA should be focused on the user itself, not others around them. For better acceptance, the ASA should have a relation to a German-speaking country (e.g. be assembled in Germany, or have a German name). ASAs should also be built with colloquial vocabulary options, if they aren't used in a strictly formal context. The social context for German speakers seems to already be more met.

Some general recommendations are unfortunately very context dependent. The physical embodiment of an ASA should be when real-life tactile interaction is present. The introversion of an ASA is entirely user-dependent. Yet, a setting for introversion may be a good option.

Additionally, ASAs should logically improve in their speed, offered functionality. This is culture-independent, and generally provides a better user experience. German speakers may benefit more from a wider service-driven functionality (instead of more social skills).

7 Conclusion and Future Work

7.1 Conclusion

The presented German version of the ASA questionnaire shows the ability to provide comparable results to the original ASA questionnaire composed in English. The construct scores especially show good to excellent correlation, with little average difference between the two languages. For the short ASA questionnaire translation, 75% of items are above

fair classification, with only 2 items being poor. In summary, the validated translation allows researchers to evaluate human-ASA interactions within a German-speaking context. With German being an official language of six countries, and often being taught as a secondary foreign language, this opens up a new realm of research possibilities.
The research allows us to answer the research question **"What are the differences and similarities of the English and German human-ASA interaction interpretations?"**. Cultural comparison shows, that German speakers rate the Agent's Personality Presence, Social Presence more positively. On the other hand, they rate more negatively the Interaction Impact on Self-Image.
To summarize future ASA development in the German-speaking context, German speakers may benefit from direct, administratively-driven ASAs which have ties to a German-speaking country. The social context matters less.

7.2 Future Work

As stated in section 4.2, due to a late launch of the questionnaire and time constraints, only 72 participants answered the first-half survey, with 82 answering the second-half one. Ideally, we require 120 participants per each half, to detect a small effect (d = .2) with an 80%+ chance.

Additionally, the codebase could be improved in legibility, and potentially conciseness. Nonetheless, efforts for making the code better commented, and legible were present throughout the study.

Lastly, future studies could pay more attention to translations. For instance, multiple alternative translations can be provided in the first formative rounds. With some luck, this may avoid the necessity of more translation rounds. The current discussions on societal issues, such as the German discussion about gender inclusivity, should also be accounted for directly in future translation and validation efforts.

8 Responsible Research

In this section, the research practices described in this paper are critically reviewed. A deeper dive is taken into the ethical aspects of the research and the reproducibility of the methods used.

8.1 Open Science Framework (OSF)

Before the any surveys mentioned in the study were conducted, OSF forms for the surveys and associated research were published. Two such forms exist for the formative [18] and summative [17] surveys. Publishing the forms in advance guarantees that the researchers have carefully thought about the procedures of the study. It also significantly reduces the possibilities of research manipulation, by clearly stating the approach.
The OSF form addresses **Self-interest Bias** and **Confirmation Bias** from Draws et al. [5]. We can not secretly change the approach to favor some result.

8.2 Data collection

A large part of the research outlined in the paper concerns data collection from online questionnaire participants. The data was not collected directly by us, but rather the supervisor. Approval from the TU Delft University Human Research Ethics Committee has been granted beforehand, ensuring a high ethical standard of the study. Before the data was passed onto us, it was anonymized by the supervisor. The supervisor removed any means of identification of individual participants, only a random identifiers were passed onto us.
The anonymization is another protection layer from **Self-interest Bias** and **Halo Effect** of Draws et al. [5].

8.3 Participant selection

To assure a high quality for the data collected, specific inclusion and exclusion criteria were created for the participants. Prolific was used as the platform for gathering participants. Participants wishing to take part in the questionnaire had to be bilingual, with German as their primary tongue and English as their fluent secondary language. Participants were sampled based on gender. For the male and female genders, an approximate 50/50 split has been created, by creating separate male and female Prolific pages. Non-binary participants were explicitly included into the female group, as Prolific's estimates showed larger availability of male participants compared to female ones.
Self-interest Bias and **Halo Effect** of Draws et al. [5] are avoided, as we do not skew data to any particular gender.

8.4 Participation reward

To encourage participation in the questionnaires, a small cash reward (the minimum allowed on Prolific) was paid to participants who have successfully completed the questionnaires. While in an ideal scenario there would be no extra motivation provided, the payment ensures a quick enrollment of people into the study. Additionally, there are 14 attention (excl. formative round 3 with 10) checks per questionnaire. Only participants who correctly answer all checks get rewarded. In this way, data from bots and poorly focusing people is filtered out from the study.
Overconfidence or Optimism Bias from Draws et al. [5] is avoided, as participants who have not estimated their abilities properly (e.g. had a poor focus) are filtered out.

8.5 Team effort

Throughout the study there has been collaboration with the following persons: N.Albers (supervisor), E.Bokel, and in smaller part K.Tessink, J.Hennsman. The persons mentioned conducted similar studies to this one. E.Bokel in particular co-researched the German translation and validation parts of this paper. For the sake of open research, E.Bokel and I have shared pre-processed data of the questionnaires. This data was later combined into one set. With two people working on the same cause, there was a more tight peer review process of the results, and thus less chance for negligence or data manipulation.
Self-interest Bias and **Confirmation Bias** from Draws et al. [5] are once more made more difficult to create, as multiple members validate parts of the study.

8.6 Reproducibility

To ensure that future researchers are able to reproduce the data provided in the paper, the codebase used in the study can be found on the 4TU.ResearchData and Github repositories. No link can be provided for the formative codebase in this paper, as at the time of writing the files are still under review. The formative assessment has DOI **10.4121/1975af9a-9b58-4dde-ae58-58e1001ef553** [16]. The summative codebase can be found with DOI **10.5281/zenodo.8079245**, see Khodakov [15].
The code has been used to conduct the analysis/evaluation. The code is a modification of the codebase from Li et al. [24]. Instructions on the inner workings of the code can be found within the files themselves. A `README` file is provided for a general overview.
Since third parties may verify the research from beginning till end, **Self-interest Bias** of Draws et al. [5] is difficult to have. Any cherry-picking would be found quickly.

8.7 Literature

To crosscheck the results from the practical part of the study, and to provide recommendations for future ASA developments, a literature study was conducted. All sources used for this paper are referenced. No cherry-picking due to confirmational bias was done. No plagiarism is present.
This adddresses **Confirmation Bias** of Draws et al. [5].

8.8 Sponsorship disclosure

This work is part of the multidisciplinary research project Perfect Fit, which is supported by several funders organized by the Netherlands Organization for Scientific Research (NWO), program Commit2Data - Big Data & Health (project number 628.011.211). Besides NWO, the funders include the Netherlands Organisation for Health Research and Development (ZonMw), Hartstichting, the Ministry of Health, Welfare and Sport (VWS), Health Holland, and the Netherlands eScience Center. The German translation is further funded by the North Rhine-Westphalia state government in Germany.

References

[1] Nele Albers, Mark A Neerincx, Kristell M Penfornis, and Willem-Paul Brinkman. Users' needs for a digital smoking cessation application and how to address them: A mixed-methods study. *PeerJ*, 10:e13824, 2022.

[2] Eileen Becks and Torben Weis. Nudging to improve human-ai symbiosis. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 132–133. IEEE, 2022.

[3] Domenic V Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4):284, 1994.

[4] Maartje MA de Graaf, Soumaya Ben Allouch, and Jan AGM Van Dijk. What makes robots social?: A user's perspective on characteristics for social human-robot interaction. In *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*, pages 184–193. Springer, 2015.

[5] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 48–59, 2021.

[6] Friederike Eyssel and Dieta Kuchenbrandt. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, 51(4):724–731, 2012.

[7] W Holmes Finch, Jocelyn E Bolin, and Ken Kelley. *Multilevel modeling using R*. Crc Press, 2019.

[8] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2022.

[9] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, and Willem-Paul Brinkman. Questionnaire items for evaluating artificial social agents-expert generated, content validated and reliability analysed. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 84–86, 2021.

[10] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. What are we measuring anyway? -a literature survey of questionnaires used in studies reported in the intelligent virtual agent conferences. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 159–161, 2019.

[11] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. The 19 unifying questionnaire constructs of artificial social agents: An iva community analysis. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

[12] Laura Hoffmann, Nikolai Bock, and Astrid M Rosenthal vd Pütten. The peculiarities of robot embodiment (emcorp-scale) development, validation and initial test of the embodiment and corporeality of artificial agents scale. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, pages 370–378, 2018.

[13] Willem KB Hofstee, Henk AL Kiers, Boele De Raad, Lewis R Goldberg, and Fritz Ostendorf. A comparison of big-five structures of personality traits in dutch, english, and german. *European Journal of Personality*, 11(1):15–31, 1997.

[14] Juliane House. Communicative styles in english and german. *European Journal of English Studies*, 10(3):249–267, 2006.

[15] Boleslav Khodakov. bolkhod/Cultural-Differences-DE-EN: Initial release with funding statement., June 2023.

[16] Boleslav Khodakov, Emma Bokel, Nele Albers, Andrea Bönsch, Jonahtan Ehret, and Willem-Paul Brinkman. German asa questionnaire translation - part 1: Translation and formative assessment. 4TU.ResearchData, June 2023.

[17] Boleslav Khodakov, Emma Bokel, Nele Albers, and Willem-Paul Brinkman. German and dutch asa questionnaire translations - part 2: Summative assessment, June 2023.

[18] Boleslav Khodakov, Emma Bokel, Kriss Tesink, Johan Hensman, Nele Albers, and Willem-Paul Brinkman. German and dutch asa questionnaire translations - part 1: Translation and formative assessment, May 2023.

[19] Jihyun Kim, Kelly Merrill Jr, and Chad Collins. Ai as a friend or assistant: The mediating role of perceived usefulness in social ai vs. functional ai. *Telematics and Informatics*, 64:101694, 2021.

[20] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[21] Brigitte Krenn, Stephanie Schreitter, and Friedrich Neubarth. Speak to me and i tell you who you are! a language-attitude study in a cultural-heritage application. *AI & society*, 32:65–77, 2017.

[22] John K Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573, 2013.

[23] A Kulhánek, R Gabrhelík, D Novák, V Burda, and H Brendryen. ehealth intervention for smoking cessation for czech tobacco smokers: pilot study of user acceptance. *Adiktologie*, 18(2):81–85, 2018.

[24] Fengxiang Li, Siska Fitrianie, Merijn Bruijnes, Amal Abdulrahman, Fu Guo, and Willem-Paul Brinkman. [under review] mandarin chinese translation of the artificial-social-agent questionnaire instrument for evaluating human-agent interaction.

[25] Jian Li, Jinsong Huang, Jiaxiang Liu, and Tianqi Zheng. Human-ai cooperation: Modes and their effects on attitudes. *Telematics and Informatics*, 73:101862, 2022.

[26] Laura Liebig, Licinia Güttel, Anna Jobin, and Christian Katzenbach. Subnational ai policy: shaping ai in a multi-level governance system. *AI & SOCIETY*, pages 1–14, 2022.

[27] Shraddha Mehta, Rowena F Bastero-Caballero, Yijun Sun, Ray Zhu, Diane K Murphy, Bhushan Hardas, and Gary Koch. Performance of intraclass correlation coefficient (icc) as a reliability index under various distributions in scale reliability studies. *Statistics in medicine*, 37(18):2734–2752, 2018.

[28] Mohammad Obaid, Maha Salem, Micheline Ziadee, Halim Boukaram, Elena Moltchanova, and Majd Sakr. Investigating effects of professional status and ethnicity in human-agent interaction. In *Proceedings of the fourth international conference on human agent interaction*, pages 179–186, 2016.

[29] Eva Ogiermann. Politeness and in-directness across cultures: A comparison of english, german, polish and russian requests. 2009.

[30] Chao Qu, Willem-Paul Brinkman, Yun Ling, Pascal Wiggers, and Ingrid Heynderickx. Human perception of a conversational virtual human: an empirical study on the effect of emotion and culture. *Virtual Reality*, 17:307–321, 2013.

[31] Maha Salem, Micheline Ziadee, and Majd Sakr. Marhaba, how may i help you? effects of politeness and culture on robot acceptance and anthropomorphization. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 74–81, 2014.

[32] Stefaan Verhulst, Andrew Young, and Mona Sloane. The ai localism canvas. *Inform Raumentwicklung*, 48(3):86–89, 2021.

[33] Sarah Theres Völkel, Ramona Schoedel, Lale Kaya, and Sven Mayer. User perceptions of extraversion in chatbots after repeated use. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[34] Jennifer Zamora. I'm sorry, dave, i'm afraid i can't do that: Chatbot perception and expectations. In *Proceedings of the 5th international conference on human agent interaction*, pages 253–260, 2017.

A Appendix

A.1 Tables

Table 1: Categories of ICC classifications by Cicchetti [3] and number of ICC values in classification category.

Classification	ICC Range	90-item set	Construct/ Dimension	24-item set
Excellent	0.75-1.00	24 (26.67 %)	18 (75%)	6 (25%)
Good	0.60-0.74	35 (38.89%)	4 (16.67%)	12 (50%)
Fair	0.40-0.59	21 (23.33%)	2 (8,33%)	4 (16.67%)
Poor	0-0.39	5 (5.56%)	0	2 (8.33%)

Table 2: ICC values and mean score differences of 24 constructs and dimensions

Construct/Dimension	"ID"	n	ICC	M - De	M - En	△ - M	△ - SD	CI - 2.5%	CI - 97.5%
Agent's Believability									
Human-Like Appearance	HLA	4	0.907	1.923	-1.163	-0.01636	0.09587	-0.2036	0.1749
Human-Like Behavior	HLB	5	0.8893	1.758	-0.3444	0.03736	0.09469	-0.1482	0.2227
Natural Appearance	NA	5	0.8317	1.498	-0.4583	0.1644	0.08831	-0.004479	0.3454
Natural Behavior	NB	3	0.9113	1.694	-0.4722	0.1372	0.08483	-0.03008	0.3027
Agent's Appearance Suita.	AAS	3	0.7615	1.329	1.236	0.1048	0.1001	-0.09632	0.297
Agent's Usability	AU	3	0.7704	1.222	1.454	-0.1297	0.09254	-0.3124	0.05009
Performance	PF	3	0.731	1.147	1.398	0.0001719	0.09211	-0.1807	0.1794
Agent's Likeability	AL	5	0.9263	1.417	0.8	0.007056	0.06101	-0.1138	0.1263
Agent's Sociability	AS	3	0.5814	1.404	0.3241	0.5203	0.137	0.2611	0.7985
Agent's Personality Presence	APP	3	0.8497	1.558	-0.5231	-0.0499	0.09854	-0.2462	0.1427
User Acceptance of the A.	UAA	3	0.7219	1.12	1.417	-0.12	0.0957	-0.3111	0.06505
Agent's Enjoyability	AE	4	0.8166	1.332	1.34	-0.1024	0.08774	-0.279	0.06596
User's Engagement	UE	3	0.5111	0.8579	1.653	0.2787	0.104	0.07542	0.4832
User's Trust	UT	3	0.7168	1.12	0.3426	0.1176	0.1029	-0.07905	0.3239
User-Agent Alliance	UAL	6	0.8288	1.048	0.4167	-0.03101	0.07408	-0.1725	0.1193
Agent's Attentiveness	AA	3	0.6597	0.9308	1.838	-0.0397	0.08686	-0.209	0.1296
Agent's Coherence	AC	4	0.8155	1.116	1.778	0.05154	0.06824	-0.0805	0.1875
Agent's Intentionality	AI	4	0.8095	1.215	0.4375	-0.1395	0.09203	-0.3202	0.04172
Attitude	AT	3	0.9177	1.422	1.449	0.05385	0.06399	-0.07057	0.1796
Social Presence	SP	3	0.8388	1.397	-0.6389	-0.002784	0.09589	-0.1914	0.1863
Interaction Impact on Self.	IIS	4	0.8732	1.21	0.2639	-0.1017	0.06486	-0.2291	0.02697
Emotional Experience									
Agent's Emotional Intellig.	AEI	5	0.8768	1.663	-0.8806	0.1018	0.08923	-0.07433	0.2733
User's Emotion Presen.	UEP	4	0.8024	1.326	0.7049	0.208	0.0878	0.03981	0.3841
User-Agent Interplay	UAI	4	0.8311	1.066	0.8993	0.0502	0.07396	-0.09643	0.1935
Grand Mean	-	-	0.7991	0.6046	0.5529	0.1069	0.0888	-	-

Note: △ Scores are pairwise differences taken from the posterior distribution. The grand mean for △ is the grand absolute mean of the mean score differences. Also, the Grand mean values are rounded to 4 decimals

Table 3: The short version of the ASA questionnaire

”ID”	Item	ICC	M - De	M - En	△ - M	△ - SD	CI - 2.5%	CI - 97.5%
HLA2	[The agent] has the appearance of a human	0.8998	-1.319	-1.306	1.77E-07	1.00E-04	-0.0001949	0.0001963
HLB5	[The agent] has a human-like manner	0.7978	-0.1667	-0.1667	-0.004375	0.1478	-0.2893	0.2903
NA4	[The agent] seems natural from the outward appearance	0.617	-0.4444	-0.5556	0.2409	0.1342	-0.02199	0.5049
NB3	[The agent] reacts like a living organism	0.7917	0.125	-0.01389	0.1607	0.1509	-0.1376	0.4545
AAS1	[The agent]’s appearance is appropriate	0.6119	1.333	1.181	0.04507	0.08933	-0.1033	0.2503
AU1	[The agent] is easy to use	0.6126	1.389	1.444	-8.67E-07	0.0001633	-0.0003239	0.0003245
PF1	[The agent] does its task well	0.7092	1.542	1.347	0.1884	0.1159	-0.04188	0.4145
AL2	I like [the agent]	0.8967	0.8472	0.75	-4.59E-07	9.76E-05	-0.0001963	0.0001868
AS1	[The agent] can easily mix socially	0.3082	0.7083	-0.3889	1.13	0.2197	0.7019	1.56
APP1	[The agent] has a distinctive character	0.5597	-0.08333	-0.2778	-0.01256	0.09579	-0.2139	0.1751
UAA1	The user will use [the agent] again in future	0.6611	1.486	1.542	6.37E-06	0.0004853	-0.0006613	0.0006497
AE1	[R] [The agent] is boring	0.8042	1	0.8056	-1.35E-07	0.0001926	-0.0003839	0.0003657
UE2	The interaction captured the user’s attention	0.5671	2	1.972	8.70E-07	0.0001245	-0.0002416	0.0002458
UT3	The user can rely on [the agent]	0.5686	0.7222	0.5694	0.1224	0.142	-0.16	0.4038
UAL1	[The agent] and the user have a strategic alliance	0.7403	0.04167	-0.1528	0.1584	0.1418	-0.1186	0.4363
AA2	[The agent] is attentive	0.3927	1.681	1.514	-0.04014	0.1142	-0.2704	0.1905
AC1	[R] [The agent]’s behavior does not make sense	0.6271	1.958	1.944	0.02148	0.1078	-0.1972	0.2481
AI3	[R] [The agent] has no clue of what it is doing	0.6752	0.9722	1.083	-0.0787	0.135	-0.3507	0.1844
AT1	The user sees the interaction with [the agent] as something positive	0.7773	1.417	1.361	-2.08E-06	0.0003503	-0.0006175	0.0005974
SP2	The agent is a social entity	0.7205	-0.7083	-0.6806	-0.008639	0.1511	-0.31	0.284
IIS2	Others would encourage the user to use [the agent]	0.7251	0.25	0.4167	2.13E-07	0.0004417	-0.0006989	0.0007248
AEI3	[R] [The agent] is emotionless	0.503	-0.1667	-0.7083	0.136	0.193	-0.212	0.5435
UEP3	The emotions the user feels during the interaction are caused by [the agent]	0.7135	1.194	1.069	0.09097	0.1004	-0.09702	0.2989
UAI4	[The agent]’s and the user’s emotions change to what they do to each other	0.6946	0.3056	0.2778	0.1063	0.1259	-0.131	0.3664
Grand Mean		0.6656	0.6701	0.5428	0.106	0.0903	-	-

Note: Codes in the items: [R] refers to a reverse-scoring questionnaire item; and [The agent] was replaced with the ASA’s name. △ Scores are pairwise differences taken from the posterior distribution. The grand mean for △ is the grand absolute mean of the mean score differences. Also, the Grand mean values are rounded to 4 decimals

Table 4: Items with credible bias indication

Item	M - German	M - English	$\triangle$ - M	$\triangle$ - SD	CI - 2.5%	CI - 97.5%	Max$\{$P ($\triangle > 0$), P($\triangle < 0$)$\}$
NA2	-0.2917	-0.5	0.2979	0.1458	0.02135	0.5886	$>$0.98
AL5	0	-0.4306	0.434	0.1208	0.2023	0.6756	$>$0.99
AS1	0.7083	-0.3889	1.132	0.2213	0.6934	1.56	$>$0.99
AS2	1.111	0.5972	0.4865	0.2025	0.08911	0.888	$>$0.99
R_APP2	-0.5	-0.1111	-0.2761	0.1423	-0.5553	-0.004548	$>$0.98
R_UAA3	0.8472	1.292	-0.3583	0.1679	-0.685	-0.02982	$>$0.98
R_AE4	1.181	1.653	-0.4268	0.1885	-0.7862	-0.04713	$>$0.98
UAL3	-0.02778	-0.4861	0.3641	0.1681	0.02752	0.6859	$>$0.98
UAL4	0.7639	1.236	-0.4468	0.135	-0.7111	-0.1793	$>$0.99
AEI1	-0.5	-1.014	0.4896	0.1785	0.1512	0.8532	$>$0.99
UAI2	1.375	0.9583	0.4287	0.1854	0.05832	0.7872	$>$0.98

Note: $\triangle$ Score are pairwise differences taken from the posterior distribution.

Table 5: Construct/dimension rating difference between mixed-international English-speaking and German primary-tongue groups

Construct/Dimension	M - German	M - English	△ - M	△ - SD	CI - 2.5%	CI - 97.5%	Max{P (△ > 0), P(△ < 0)}
Agent's Believability							
HLA	-1.163	-0.7533	-0.3098	0.1631	-0.6287	0.0107	0.9728
HLB	0.04398	-0.3444	0.3374	0.1732	-0.00315	0.676	0.9749
NA	-0.4583	-0.2429	-0.1788	0.1513	-0.4744	0.1173	0.8759
NB	-0.2932	-0.4722	0.132	0.1616	-0.1837	0.449	0.7918
AAS	1.236	1.346	-0.1149	0.1415	-0.392	0.1629	0.7957
AU	1.234	1.454	-0.2016	0.1344	-0.4635	0.06166	0.9309
PF	1.398	1.306	0.08222	0.1315	-0.1758	0.3415	0.7296
AL	0.7699	0.8	-0.01383	0.1501	-0.3091	0.2787	0.5397
AS	0.3241	0.3164	0.004234	0.1654	-0.3191	0.3293	0.5101
APP	0.1986	-0.5231	0.6846	0.1601	0.3702	0.9979	1
UAA	1.417	1.311	0.08802	0.1347	-0.1752	0.3527	0.7492
AE	1.252	1.34	-0.06004	0.136	-0.3282	0.2056	0.6687
UE	1.653	1.812	-0.1735	0.1213	-0.4102	0.06466	0.9283
UT	0.4311	0.3426	0.1103	0.1383	-0.1599	0.3802	0.7864
UAL	0.4167	0.5125	-0.1025	0.136	-0.3694	0.1651	0.784
AA	1.654	1.838	-0.1582	0.138	-0.4274	0.1139	0.8758
AC	1.778	1.549	0.2048	0.1269	-0.0442	0.4534	0.9488
AI	0.6852	0.4375	0.2735	0.1484	-0.01626	0.5649	0.9696
AT	1.449	1.431	-0.0164	0.1411	-0.2918	0.26	0.5536
SP	-0.1629	-0.6389	0.4338	0.1747	0.09099	0.777	0.9946
IIS	0.2639	0.648	-0.3871	0.1348	-0.6514	-0.1222	0.9977
Emotional Experience							
AEI	-0.6684	-0.8806	0.1403	0.1744	-0.2009	0.4836	0.7878
UEP	0.7049	0.6245	0.1101	0.1413	-0.1677	0.3869	0.7877
UAI	0.7946	0.8993	-0.1216	0.1376	-0.3922	0.1474	0.8084

Note: △ Score are pairwise differences between German primary-tongue cultural background and mix-international cultural background taken from the posterior distribution.

A.2 Participant statistic of Prolific studies

The following are statistics for successful Prolific participants whose submissions are the basis of this study:

- Formative round 1:
  1. Amount of participants who identified as male: 30. Expressed as percent: 50%
  2. Amount of participants who identified as female: 27. Expressed as percent: 45%
  3. Amount of participants who identified as non-binary: 3. Expressed as percent: 5%
  4. Age range of participants: 19 - 73
  5. Mean age (rounded): 35
  6. Standard Deviation of age (rounded): 13
- Formative round 2:
  1. Amount of participants who identified as male: 15. Expressed as percent: 50%
  2. Amount of participants who identified as female: 13. Expressed as percent: 43.3%
  3. Amount of participants who identified as non-binary: 2. Expressed as percent: 6.7%
  4. Age range of participants: 22 - 70
  5. Mean age (rounded): 35
  6. Standard Deviation of age (rounded): 12
- Formative round 3:
  1. Amount of participants who identified as male: 15. Expressed as percent: 50%
  2. Amount of participants who identified as female: 15. Expressed as percent: 50%
  3. Amount of participants who identified as non-binary: 0. Expressed as percent: 0%
  4. Age range of participants: 21 - 46
  5. Mean age (rounded): 31
  6. Standard Deviation of age (rounded): 6
- Summative round. 72 participants from first half, 82 participants from second half (we did not remove 10 participants here):
  1. Amount of participants who identified as male: 82. Expressed as percent: 53.2%
  2. Amount of participants who identified as female: 69. Expressed as percent: 44.8%
  3. Amount of participants who identified as non-binary: 3. Expressed as percent: 1.9%
  4. Age range of participants: 19 - 69
  5. Mean age (rounded): 31
  6. Standard Deviation of age (rounded): 10
  7. Participants who recommended using their data: 141

A.3 Contribution sheet

Boleslav Khodakov (German group)

- co-created formative OSF form
- co-created summative OSF form
- created transformation code for first formative round
- created transformation code for second formative round
- created transformation code for third formative round
- created evaluation code for first formative round
- created evaluation code for second formative round
- created evaluation code for third formative round
- created legend files for formative rounds
- created readme files for formative rounds
- created prolific statistics code for formative rounds
- created equalization code for summative round (with German-English data in mind)
- created culture-data creation code for summative round (with German-English data in mind)
- created transformation code for summative round (with German-English data in mind)
- created evaluation code for summative round (with German-English data in mind)
- created legend files for summative round
- created readme files for summative round (German-English version)
- created prolific statistics code for summative round
- Set up the first half of the Prolific study (round 1)
- prepared Qualtrics survey for first half of first round
- tested all Qualtrics surveys for bugs
- Created dummy data for questionnaires
- Co-created Excel documents to send to the translators for formative rounds

Emma Bokel (German group)

- co-created formative OSF form
- co-created summative OSF form
- Set up the second half of the Prolific study
- Created all Qualtics questionnaires except the first round, first half
- Created dummy data for said questionnaires
- Triple checked the questionnaires to make sure the labels were all correct
- Started a python script to calculate ICC values in the first round, but Bolek figured out how to run the R code first, so this was never completed or used
- Adjusted Bolek's code to work for the first round, second half
- Helped transform the data in the analysis code
- Created Excel documents to send to the translators
- Created the full document of the translated ASA questionnaire in German

Kriss Tesink (Dutch group)

- Co-created formative OSF form
- Co-created Qualtrics rounds 1 and 2
- Assisted in creation of dummy data for Qualtrics questionnaires
- Tested Qualtrics rounds 1 and 2 for bugs
- Created prolific codes
- Created Excel documents to be sent to translators
- Created comments for R code for round 1 and 2
- Created evaluation code for round 2
- Wrote R code to find the best alternative translations in round 2
- Assisted in evaluation R code for round 1
- Created the full document of the translated ASA questionnaire in Dutch
- Created code for summative assessment

Johan Hensman (Dutch group)

- Co-created formative OSF form
- Co-created Qualtrics rounds 1 and 2
- Created dummy data for the Qualtrics questionnaires
- Created evaluation code for round 1
- Assisted in evaluation code for round 2
- Tested Qualtrics first half of round 1 and round 2 for bugs
- Created legend files for the translation rounds
- Created Readme files for the translation rounds
- Assisted in the Excel files that were sent to the translators
- Created data transformation code for summative assessment (for Dutch-Chinese version)
- Created evaluation code for summative assessment (for Dutch-Chinese version)
- Created legend files for the summative assessment (for Dutch-Chinese version)
- Created Readme files for the summative assessment (for Dutch-Chinese version)
- Provided comments for code for the summative assessment (for Dutch-Chinese version)