NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Anton Angelgardt, Elena S. Gorbunova,

Maria Chumakova

AN ASSESSMENT OF TRUST IN ARTIFICIAL INTELLIGENT AGENTS: TOOL DEVELOPMENT

BASIC RESEARCH PROGRAM
WORKING PAPERS

SERIES: PSYCHOLOGY

WP BRP 128/PSY/2021

This Working Paper is an output of a research project implemented within NRU HSE's Annual Thematic Plan for Basic and Applied Research. Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.

Anton Angelgardt,[1] *Elena S. Gorbunova,*[2]

Maria Chumakova[3]

AN ASSESSMENT OF TRUST IN ARTIFICIAL INTELLIGENT AGENTS: TOOL DEVELOPMENT

The current state of technology means people find themselves interacting with and having to trust artificial intelligent agents. However, despite the considerable history of trust studies, there is no agreement on what trust is and what this construct consists of.

The study elaborates on the construct of trust in artificial intelligent agents and develops a questionnaire to assess this construct. Reliability, validity, internal consistency, and other essential statistical parameters of the scale are examined. In addition, difficulty and discrimination coefficients analysed to measure their properties. Confirmatory factor analysis is used to verify the theoretical structure of the developed construct on empirical data. As a result, the conclusion about the applicability of the developed scale is made.

JEL Classification: Z.

Keywords: trust, artificial intelligence, questionnaire development.

[1] HSE University. Laboratory for Cognitive Psychology of Digital Interfaces User. Research Assistant; E-mail: aangelgardt@hse.ru

[2] HSE University. School of Psychology, Laboratory for Cognitive Psychology of Digital Interfaces User. Associate Professor, Laboratory Head; E-mail: esgorbunova@hse.ru

[3] HSE University. Faculty of Social Sciences, School of Psychology. Associate Professor; E-mail: mchumakova@hse.ru

Introduction

The Internet is not a new technology. It was invented in the 1960s (Hauben & Hauben, 1997); however, in the 1990–2000s, it spread around the world at extraordinary speed (Castells, 2014). Since its invention, Internet technologies have noticeably changed how researchers study them. In classical engineering psychology, the subject-object paradigm was predominant. According to this paradigm, human-automation interaction was considered the control of a technical object by a person. Therefore, a goal of engineering psychology was defined as ensuring the reliable and efficient work of an operator in an automated control system (Zinchenko & Panov, 1962). This required the development of information models that focused on the functional characteristics and capabilities of the operator (Zinchenko, Leontiev, Panov, 1964). Later, as technologies became more complex, they acquired properties that equalized the interaction between human and machine (e.g., purposefulness, natural language use, unpredictability, etc.). One of the most famous and common approaches that overcome a classical paradigm is activity-centered design (Norman, 2013). This approach grows from Activity Theory (Leont'ev, 1978) and focuses on the users' tasks and goals rather than attempting to accommodate the users. This change of frame reflects the shifting of the researchers' views on the human-machine interaction to a subject-subject interaction. Within interpersonal relationships, the critical point is trust between two (or more) actors, particularly in a context of limited information (Borum, 2010). Therefore, the issue of trust in the use of technical equipment is crucial for modern research.

Current technology has opened an intriguing area of psychological research. In the age of the Internet of Everything, Machine Learning, and Big Data, humans have to interact with a new type of agent called intelligent agents. Trust mediates between the reliability of autonomy and the operator's ability to collaborate with intelligent agents effectively (Lee & See, 2004, Lyons & Havig, 2014).

The modern IT industry and business are interested in studies of trust in artificial intelligence (AI) because the future of these technologies depends on trust, as people are interested in transparent algorithms and interpretable outputs (Rao & Cameron, 2018). In the IT industry, trust in AI is a widely debatable theme and many businesses are experimenting with AI technologies (KPMG LLP, 2018). However, studies attempting to measure trust use unstandardized measuring methods as there is no tool for measuring trust in AI agents.

This study develops a scale of trust in AI agents. Within this broad purpose, the research focuses on creating a questionnaire that requires item generation, qualitative and quantitative approbation, and the validation and assessment of the statistical properties of the scale. Since there are no other valid

and reliable scales to measure such a construct, conducting convergent validation is problematic. Therefore, the validation procedure emphasizes different kinds of validity, such as criterion validity.

Construct elaboration

Conceptualization of trust

The study of trust has a long history. Initially, economics and philosophy focused on how people developed trust in each other. From this perspective, trust was seen as an interpersonal relationship based upon business experience (Uslaner, 2018). Then, since Rosenberg examined trust in terms of putting faith in strangers (Rosenberg, 1956), the study of psychological and social dimensions of trust had started (Kramer & Isen, 1994).

One of the first generalization of trust in psychology was giver by Baier (1986). He asserts that trust "is accepted vulnerability to another's possible but not expected ill will toward one". Later, Hardin (2006) made the next step and considered "trust as encapsulated interest". This work assumes that trust represents that the trustee takes the truster's interests correctly, so trust shows the perception of others as trustworthy. Additionally, Hardin formulated a three-component model, which includes a truster (who trusts), a trustee (who is trusted), and a behavior (a specific activity that a truster expects from a trustee). The model also describes the mechanisms of encapsulating interests. These mechanisms are based on such concepts as love and friendship, valuable relationships, and general reputation. Then, Bauer (2019) extends Hardin's model by adding two essential components, the situational context and time, defining trust as time and context specific. Other studies focus on the stability of trust (Paxton & Ressler, 2018). They point out that trust is a risk estimation while we are dealing with other people and that a good experience might lead to a higher level of trust. This finding allows us to conduct criterion validation.

Many different approaches and models emphasize various aspects of trust. However, most researchers generally agree that trust designates an expectation, and the general concept of trust may be formulated in terms of probability. This research uses Bauer's definition (2019) of trust as the "truster's subjective estimate of the probability that the trustee will display truster's preferred behavior".

Artificial Intelligent Agents

Approaches to defining AI usually are grouped into four categories (Russell & Norvig, 1995). The first one is "Thinking Humanly," which is defined as "activities that we associate with human thinking, activities such as decision-making, problem-solving, learning" (Bellman, 1978). The second one is "Thinking Rationally," which may be described as "the study of the computations that make it possible to perceive, reason, and act" (Salin & Winston, 1992). The third and fourth categories contain the "Acting" approaches. "The study of how to make computers do things at which, at the moment, people are better" (Rich and Knight, 1991) is "Acting Humanly" while "Computational Intelligence is the study of the design of intelligent agents" (Poole et al., 1998) is "Acting Rationally." Although each approach underlines various aspects of AI, the 'agent' concept is similar for all. An agent is anything that perceives its environment and acts upon that environment (Russell & Norvig, 1995). Therefore, intelligent systems (or agents) are technologically advanced machines that perceive and respond to the world around them. They can take many forms, from automated vacuums or YouTube recommendations to facial recognition programs or Amazon's personalized shopping suggestions.

Trust in Artificial Intelligence

While the concept of trust is more or less explicated, trust in AI is a relatively new one. In several studies on trust in technical equipment (Akimova & Oboznov, 2016, Nestik, 2018), trust has a structure consistent with a classical attitude structure proposed by Rosenberg and Hovland (1960). This leads to a misunderstanding of the concept because, as described above, an essential aspect of trust is an expectation of specific behavior. Regarding this Faulkner and Simpson (2017) comment that "trust may be conceptualized as a cognitive attitude or belief, and while being so, trust is neither a conative attitude, or intention, nor an affective attitude, or emotion".

In the area of human-automation interaction research, there are models of the human-machine teams. For example, Pynadath et al. (2018) proposed a model closest to the psychological point of view which grows from the concept of situational awareness (SA), and includes information about the partners' situations, their plans, and future implications and uncertainties. The SA-based Agent Transparency (SAT) defines the essential information that a human-automation team must share for effective collaboration (Chen et al., 2018). Moreover, the required agent transparency level leads to appropriate trust and gives the operators the ability to calibrate their interactions and minimize disuse and misuse. This model shows that knowledge about the principles of a system structure and its working is necessary for its practical usage.

The most elaborated model of trust in AI was presented by Larsen (2018). The model combines old and modern trust research results and defines six dimensions. *Predictability* is the truster's subjective assessment of the trustee's trustworthiness; *consistency* is the extrapolation of prior behavior to posterior expectations; *utility* is the cost and importance of trust; *faith* is an act of accepting a context outside the boundaries of what is known; *dependability* is the willingness to place oneself as a truster in a position of risk; and *understanding* is the knowledge of how the system works. Therefore, the model may be used as the basis for developing a scale of trust in AI agents.

Construct operationalization

We chose Larsen's model as a theoretical framework for developing a psychometric construct of trust in artificial intelligent agents (TAIA). Six attributes of TAIA are latent variables. Each attribute measures a part of TAIA; therefore, TAIA is a second-level factor.

The factors were operationalized as:

- Predictability (PR)
    - Participants assess a system as trustworthy
    - Participants consider the result of system work as predictable
- Consistency (CO)
    - Participants can extrapolate prior experience of dealing with systems onto posterior expectations
    - Participants can estimate the trustworthiness of a system based on their experience of using it
- Utility (UT)
    - Participants believe that intelligent technologies are valuable for society
    - Participants think that it is important to trust AI systems
    - Participants consider AI technologies as useful
- Faith (FA)
    - Participant are ready to trust the system even they do not know how it works
    - Participants accept a context outside the boundaries of what is known
- Dependability (DE)
    - Participants are ready to place themselves in a position of risk
- Understanding (UN)
    - Participants understand how an algorithm of a system works
    - Participants understand how to interact with a system

Pre-test items generation

Based on the operationalization, the authors generated 62 pre-test items. The following areas of content were taken into consideration (Bisen, 2019):

- Virtual assistants and chatbots
- E-commerce
- Self-driving cars and autonomous vehicles
- Security and surveillance
- Financial analytics
- Educational AI
- Healthcare and Medical Imaging Analysis

Validation

The General Trust Scale (GTS) (Yamagishi & Yamagishi, 1994) was used to validate the scale since GTS is more general than TAIA. The hypothesis of the GTS and TAIA correlation was that the two scales are slightly positively correlated. The expected value of Pearson's correlation coefficient is 0.2–0.3. However, an alternative hypothesis was also formulated: GTS and TAIA scores are slightly negatively correlated. An alternative hypothesis appeared because a juxtaposition 'human vs. AI' exists, and the GTS was developed for estimating trust in humans.

The criterion validation arises from previous research (Paxton & Ressler, 2018). As mentioned above, a good experience usually leads to a higher level of trust. Therefore, self-reported scores from past dealings with AI technologies are expected to correlate with TAIA scores positively.

Qualitative approbation of the items

Methods

A think-aloud interviewing technique was used to conduct qualitative approbationconfirmation, as it is an open-ended format of the cognitive interview. This type is appropriate to examine how the construct of interest is represented in participants' consciousness. Such an interview format requires less interview training and is free from interviewer-imposed bias. Cognitive interviews were conducted based on Willis' guide (2005).

Participants

Ten volunteers participated in the study. They were recruited with the snowball technique from various cities and social strata. The average age of subjects was 41 years (from 19 to 68). There were four males and six females in a sample.

Data collection

A set of pre-test items, grouped by latent factors, was presented to participants. Subjects were explicitly instructed to "think aloud" as they answered the survey questions. Respondents should rate their agreement with statements on a seven-point scale (‘strongly disagree’ to ‘strongly agree’). An interviewer recorded the subjects’ reasoning and comments. When needed, the researcher asked clarifying questions using verbal probing techniques (interpretation probe, paraphrasing, confidence judgment, etc.).

All participants successfully managed a task after simple training. Data about item complexity, ambiguity, and statement structure was collected. This data identified significant issues, and then it data collection stopped.

Data analysis

Data from each participant was summarized on a question-by-question basis, and then findings were further aggregated over the interview to obtain a complete review of the questionnaire draft. The analysis detected dominant trends (repeatedly emerging issues) across interviews and so-called “discoveries” that are unique problems that may affect data quality in the actual survey.

Results and discussion

No statements had critical problems, and none of them were excluded. However, seven questions were complex and had to be simplified. Another seven questions were ambiguous; therefore, these items were clarified and rephrased into a more precise form. Some items had unique problems such as mixing similar aspects of the concept, difficulty, or were too broad.

The primary trend was insufficient experience of dealing with AI agents. This was expected since older people and people from small Russian towns do not widely use such technologies. With this discovery, the authors assumed that a Russian sample might not reveal the theoretical factor of consistency. This assumption refers to the lack of experience of dealing with AI technologies which is crucial to extrapolate prior behavior of systems onto posterior expectations. Another factor that is influenced by the rarity of AI usage is understanding. Most of the participants noted that they do not

understand how AI agents work. Hence, they consider items of understanding subscale as statements about the interaction between a human and an intelligent system rather than the internal working of the system.

Generally, respondents could adequately describe what items were asking about and why they were grouped so. Therefore, this result may be considered as evidence of the face validity of the questionnaire.

Quantitative approbation of the scale

Methods

Participants and data collection

Participants were recruited via Yandex.Toloka platform. The task was available only for Russian residents. Additionally, task availability was restricted to the top 20% of users to raise data quality.

After opening the task on Yandex.Toloka, subjects had to click on a button redirecting them to the OneClick Survey platform containing the survey. The survey consists of an instruction page, a page with definitions of used terms, TAIA items, GTS items, questions aimed at scale validation, and a demographic information form. The page with definitions was intended to familiarize respondents with terms that are used in a questionnaire (e.g., “artificial intelligence,” “smart home,” “virtual assistant,” etc.). The TAIA scale was shown on a single page. Items were grouped in six blocks by subscales. Block sequence and the sequence of the items in each block were randomly mixed for each participant. We added to the TAIA item set 15 additional items, equally distributed per block, to control the quality of questionnaire responses. Six items of the GTS were presented on a single page with a fixed sequence. Questions about the experience of dealing with AI technologies featured if-else conditions to allow participants to skip the questions if they had not used some technologies (e.g., self-driving cars). Respondents evaluated their experience of dealing with intelligent systems on a six-point scale (‘extremely negative’ to ‘extremely positive’). The demographic form included fields for age, gender, current city, level and specialization of education, area of work, and job position. A reward for completing the questionnaire was $US0.7.

A total of 620 volunteers participated in the study, and 513 of them completed the survey. Eighteen participants were excluded from the analysis since they made more than five mistakes in the additional items. There were 233 (47%) females and 262 (53%) males in a final sample. The average age of females was 35.5 (SD = 11.1) while for males it was 36.0 (SD = 10.1). Gender groups were equivalent by age ($t(472.19) = -0.70$, $p = .49$). Twenty-two per cent of subjects live in Moscow and

Saint Petersburg, and the rest of the sample was from a wide variety of Russian cities. Participants were educated in various specialties, e.g., economics, management, law, medicine, engineering, psychology, IT, art, design, HR, etc. About 13% of the sample had a second specialization.

Data analysis

Data analysis was conducted with R version 4.0.3 (R Core Team, 2020). Psychometric analysis was carried out with "psych" package version 2.0.9 (Revelle, 2020). We assessed item and scale quality based on Classical Test Theory (Allen, Yen, 2002). As a measure of internal consistency, Cronbach's α coefficient was calculated. A split-half method was used with Guttman's λ-6 and average split-half reliability metrics to assess subscale reliability.

To verify the theoretical construct structure on the empirical data, we conducted confirmatory factor analysis using the "lavaan" package version 0.6-8 (Rosseel, 2012). Validation was executed by correlation analysis using Pearson's correlation coefficient.

Data was wrangled with the "tidyverse" package version 1.3.0 (Wickham, 2019).

Results

Score distribution

Score distributions are shown in Figure 1. For most items, skewness and kurtosis are close to or slightly differ from zero, evidencing good item wording. Most negatively skewed distributions are found for the Utility subscale which may represent a sample bias or the content of the latent factor since some questions might be sensitive. Some items have positive skewness which may be caused by wording.

Fig. 1. Score distributions of each item in subscales.

Correlations

Correlation analysis was conducted to clarify the relations of items in the subscales. The results are shown in Figure 2. Based on correlograms, it may be supposed that some items have negative discrimination (e.g., co07, ut10, and de04). In addition, several items have low correlation coefficient values, and this may result in low factor loadings in the factor analysis.

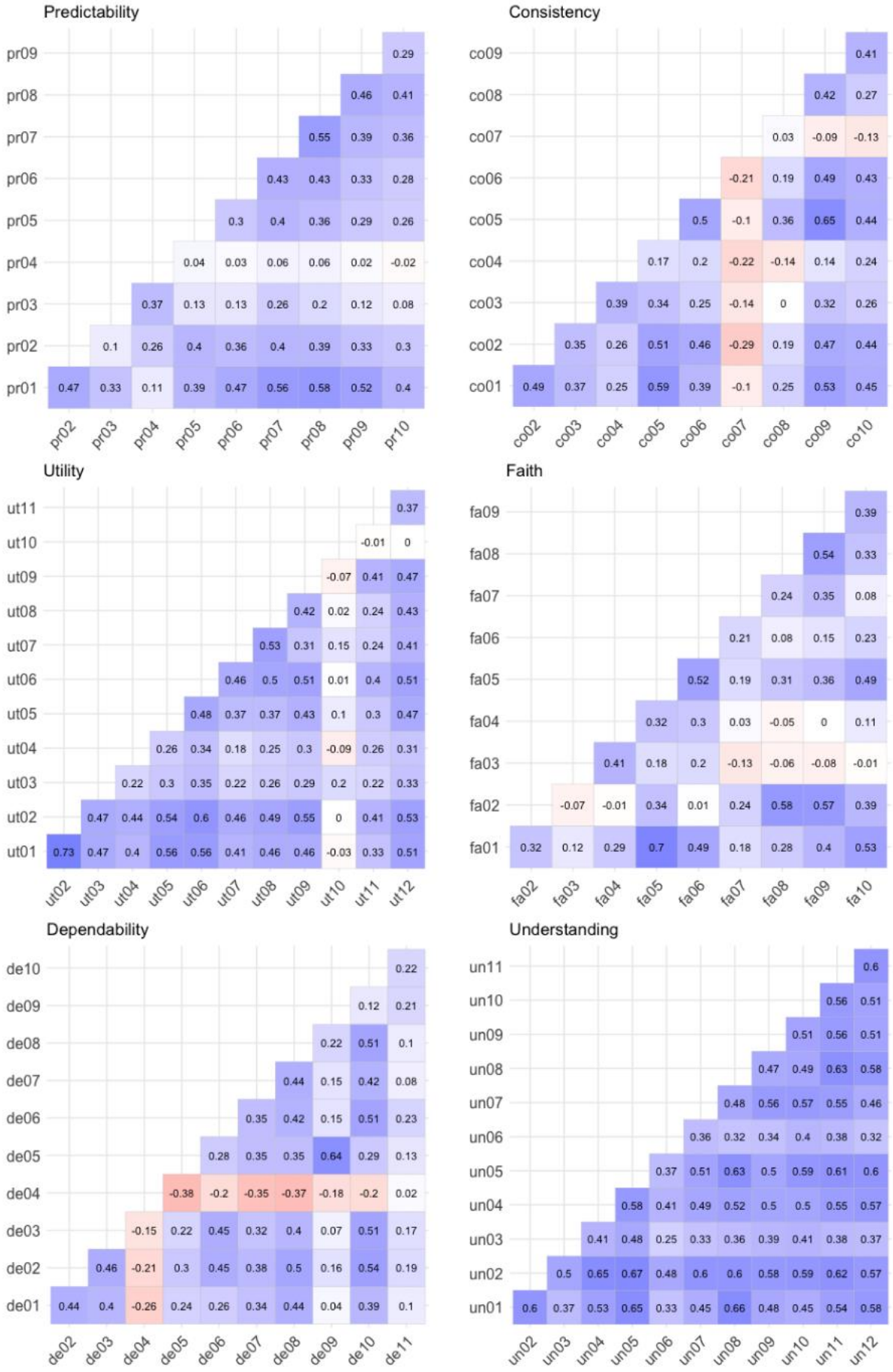

Fig. 2. Interitem correlations in each subscale.

Psychometric analysis

The values of internal consistency and split-half reliability measures are presented in Table 1. All values are greater than conventional thresholds. Item characteristics are shown in Figure 3.

Tab 1. Internal consistency and split-half reliability of TAIA subscales (before item exclusion)

Subscale	Cronbach's α	Guttman's λ-6	Average split-half reliability
Predictability	0.81	0.82	0.81
Consistency	0.77	0.80	0.77
Utility	0.86	0.87	0.86
Faith	0.77	0.81	0.77
Dependability	0.75	0.80	0.74
Understanding	0.92	0.92	0.92

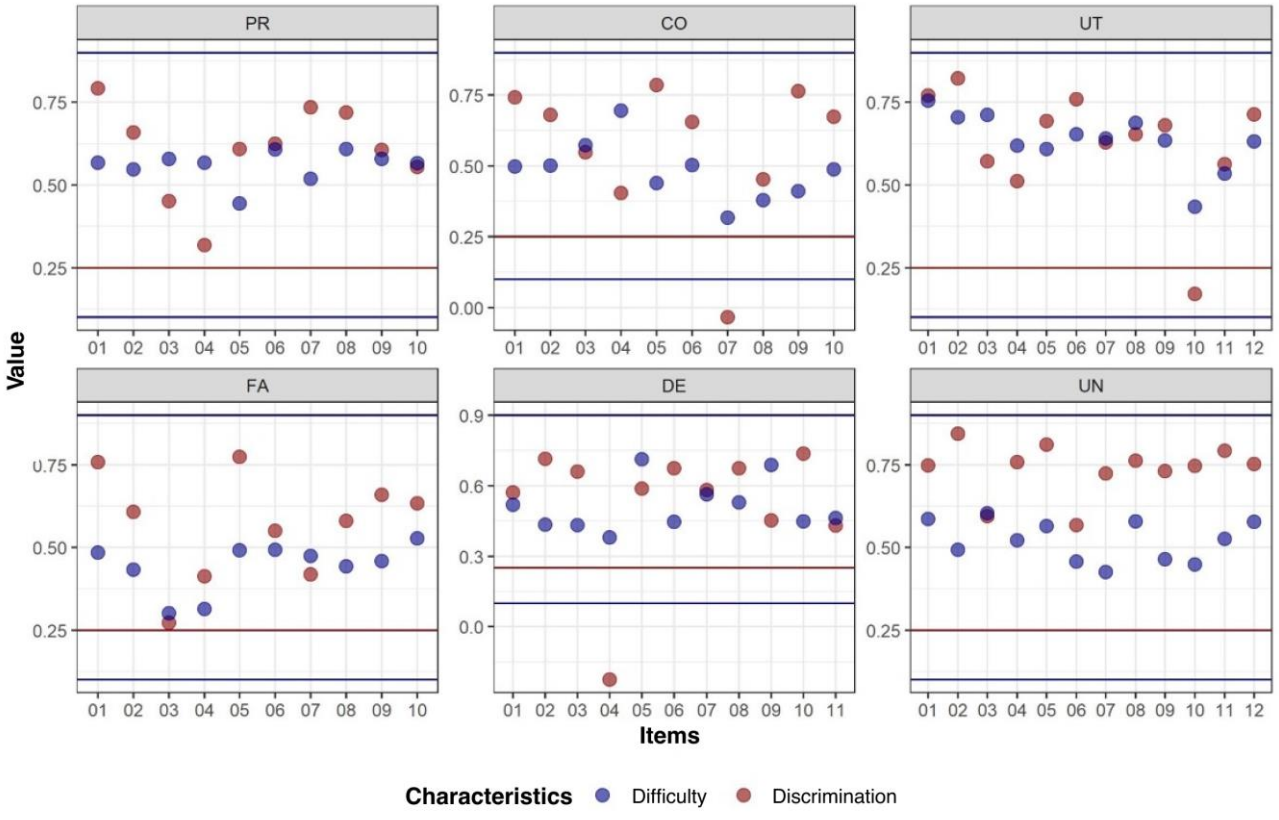

Fig. 3. Item characteristics (difficulty and discrimination).

As expected, there are three items (co07, ut10, de04) with negative discrimination. We discuss the reasons in the section below. After these items were excluded, the reliability of consistency, utility,

and dependability scales increased. All subscales have satisfactory values of internal consistency (greater than 0.75). The internal consistency and split-half reliability values after low-quality items exclusion are presented in Table 2.

Tab 2. Internal consistency and split-half reliability of TAIA subscales (after item exclusion)

Subscale	Cronbach's α	Guttman's λ-6	Average split-half reliability
Predictability	0.81	0.82	0.81
Consistency	0.82	0.83	0.82
Utility	0.88	0.88	0.87
Faith	0.77	0.81	0.77
Dependability	0.82	0.84	0.82
Understanding	0.92	0.92	0.92

Confirmatory Factor Analysis

Confirmatory factor analysis was conducted to assess whether the empirical construct structure is consistent with the theoretical model. Firstly, we tested a basic factor structure. The values of model fit measures are shown in Table 3.

Tab. 3. Model fit quality (basic model)

χ^2	df	p	AGFI	CFI	TLI	SRMR	RMSEA
6326.22	1814.00	< .001	0.610	0.710	0.697	0.101	0.071

All factor variances are significantly different from zero (Tab. 4). Although all items are correlated with the latent variables ($p < .001$ for all regression coefficients), there are several items with low factor loadings (Tab. 5).

Tab. 4. Factor variances (basic model)

Factor	Estimate	Standard Error	z	p
PR	0.605	0.059	10.207	< .001
CO	0.633	0.069	9.144	< .001
UT	0.659	0.066	9.942	< .001
FA	0.816	0.077	10.599	< .001
DE	0.377	0.058	6.490	< .001
UN	0.604	0.064	9.418	< .001

Tab. 5. Factor loadings (basic model)

Item	PR	CO	UT	FA	DE	UN
01	0.79	0.74	0.77	0.82	0.56	0.74
02	0.58	0.69	0.84	0.41	0.70	0.83
03	0.30	0.50	0.50	0.17	0.61	0.54
04	0.11	0.34	0.47	0.35		0.73
05	0.57	0.78	0.65	0.84	0.52	0.81
06	0.59	0.62	0.75	0.60	0.64	0.49
07	0.73		0.57	0.28	0.60	0.69
08	0.73	0.34	0.62	0.39	0.69	0.75
09	0.58	0.73	0.66	0.48	0.31	0.69
10	0.50	0.60		0.59	0.72	0.71
11			0.52		0.24	0.77
12			0.69			0.74

Notes: Values in the first column represents the number of items in the subscale. If the subscale does not contain an item with such a number, the corresponding cell is empty.

Since the research aims to develop a tool applicable to industrial research, the authors excluded items with loadings lower than 0.4. For the same reason, the authors tested a model with a second-level factor of TAIA. The final model is shown in Figure 4, and the measures of its fit are presented in Table 6. Factor loadings are in Table 7.

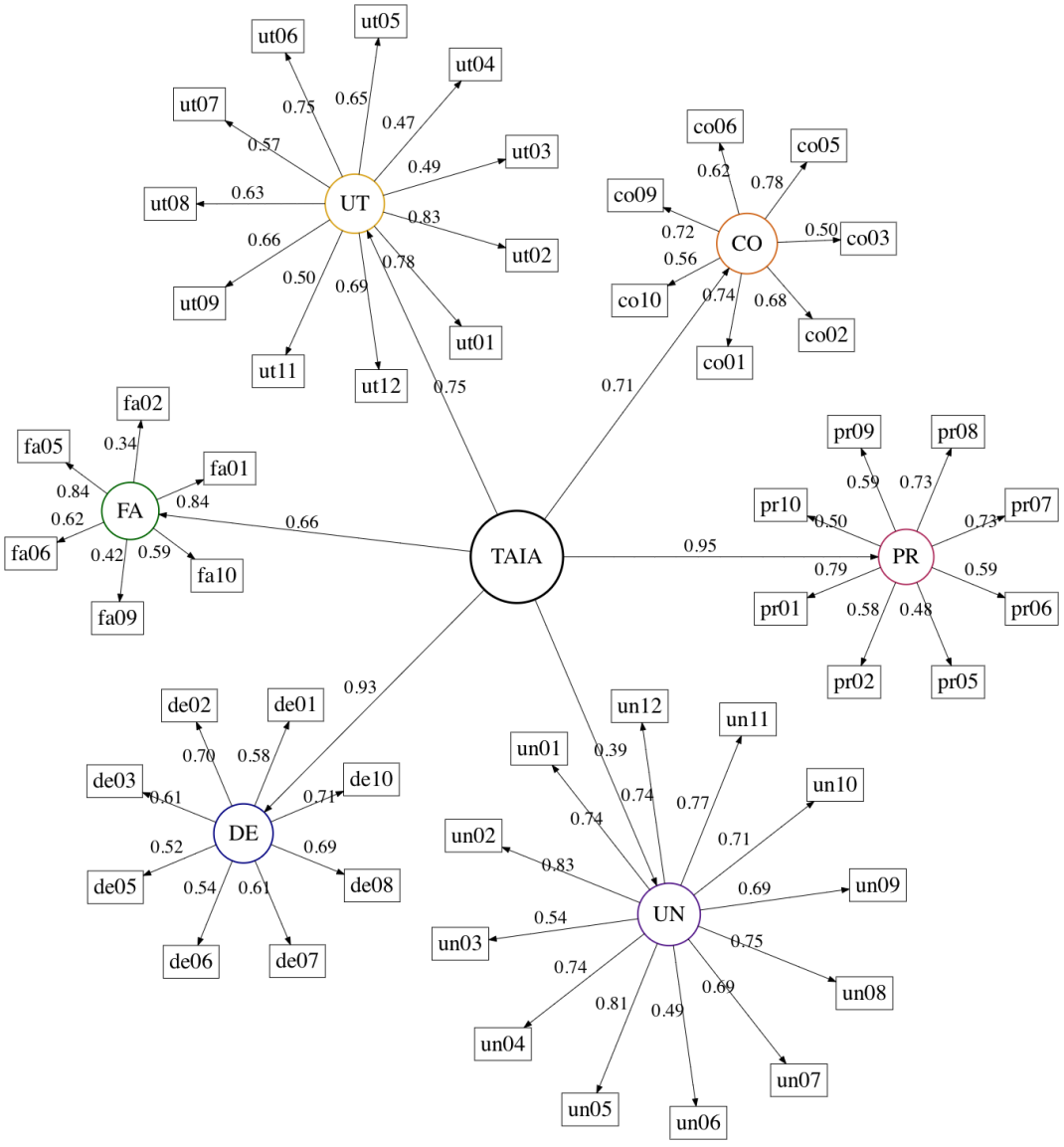

Fig. 4. Final model.

Tab. 6. Model fit quality (final model)

χ^2	df	p	AGFI	CFI	TLI	SRMR	RMSEA
3364.98	1264.00	< .001	0.748	0.841	0.833	0.085	0.058

Tab. 7. Factor loadings (final model)

Item	PR	CO	UT	FA	DE	UN	TAIA
01	0.79	0.74	0.77	0.84	0.58	0.74	
02	0.58	0.68	0.83	0.33	0.70	0.83	
03		0.50	0.49		0.61	0.54	
04			0.47			0.73	
05	0.48	0.78	0.65	0.84	0.52	0.81	
06	0.59	0.62	0.75	0.62	0.54	0.49	
07	0.73		0.57		0.61	0.69	
08	0.73		0.63		0.69	0.75	
09	0.59	0.72	0.66	0.42		0.69	
10	0.50	0.60		0.59	0.71	0.71	
11			0.50			0.77	
12			0.69			0.74	
PR							0.95
CO							0.71
UT							0.75
FA							0.66
DE							0.93
UN							0.39

Notes: Values in the first column represent the number of items in the subscale. If the subscale does not contain an item with such a number, the corresponding cell is empty.

Total scores

The fitted values from the CFA model contain a precise measure of the subscale scores and the total level of trust. However, as industrial researchers are interested in simple calculations, a more practical option is to sum item raw scores. We examined correlations between these two metrics to evaluate how direct sums capture the fitted value patterns. All coefficients are greater than 0.9 (the lowest is 0.92 and the highest is 0.99); direct sums of items scores fully represent fitted value patterns and may be used for further validation.

Scale validity

Convergent validity was examined by correlating the total and subscale scores with the GTS score. The total score and all subscales, except faith, are slightly but significantly correlated with GTS (Fig. 5). Correlation coefficients are shown in Table 8.

Tab. 8. Correlations between TAIA and GTS

	r	t	df	p
PR	0.16	3.51	493	< .001
CO	0.16	3.56	493	< .001
UT	0.11	2.47	493	.014
FA	0.08	1.87	493	0.062
DE	0.20	4.47	493	< .001
UN	0.15	3.56	493	< .001
TAIA	0.20	4.58	493	< .001

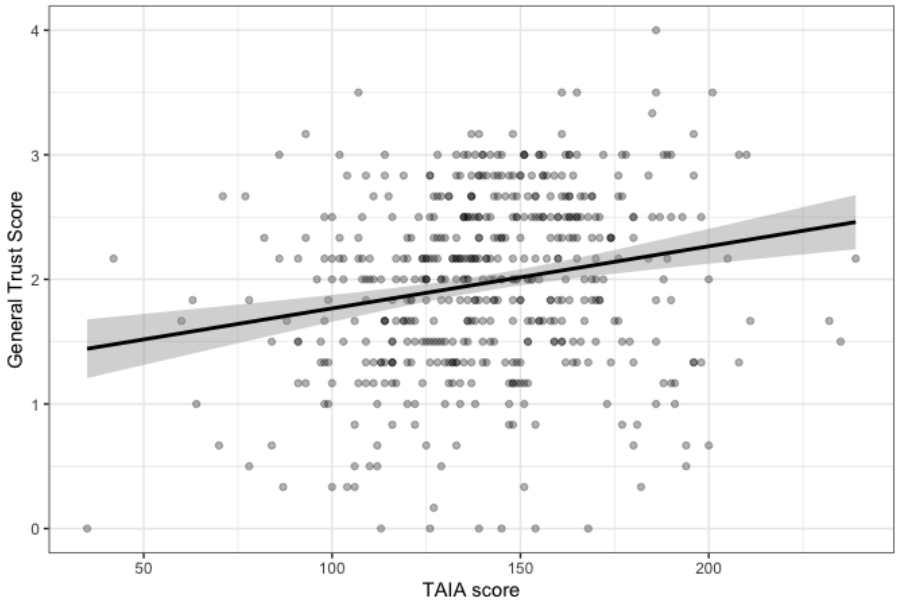

Fig. 5. Correlation between GTS score and TAIA total score.

Criterion validation was conducted by studying relations between subscale scores and experience dealing with digital technologies, as described above. Only the understanding subscale shows no correlation with experience dealing with digital technologies (Tab. 9–11).

Tab. 9. Correlations between TAIA Scale and experience of dealing with Digital Assistants

	r	t	df	p
PR	0.33	6.46	335	< .001
CO	0.23	4.29	335	< .001
UT	0.29	5.51	335	< .001
FA	0.21	3.93	335	< .001
DE	0.27	5.09	335	< .001
UN	0.09	1.71	335	.088
TAIA	0.32	6.19	335	< .001

Tab. 10. Correlations between TAIA Scale and experience of dealing with Recommender Systems

	r	t	df	p
PR	0.22	4.62	433	< .001
CO	0.27	5.72	433	< .001
UT	0.19	4.09	433	< .001
FA	0.15	3.20	433	0.001
DE	0.25	5.42	433	< .001
UN	0.09	1.79	433	0.075
TAIA	0.25	5.43	433	< .001

Tab. 11. Correlations between TAIA Scale and experience of dealing with Internet Technologies in Education

	r	t	df	p
PR	0.48	5.74	110	< .001
CO	0.44	5.17	110	< .001
UT	0.37	4.19	110	< .001
FA	0.33	3.71	110	< .001
DE	0.52	6.37	110	< .001
UN	0.30	3.31	110	.001
TAIA	0.53	6.62	110	< .001

Discussion

The first goal of the quantitative approbation was to evaluate the psychometric characteristics of items. Some items showed negative discrimination. One reason for that is the direction of the items. We revised all these items with content analysis, and the procedure revealed that participants might interpret all of those in both directions. For that reason, these items were excluded from the scale. All other items have satisfactory values of psychometrics characteristics.

The second goal was to assess the psychometric characteristics of the subscales. After excluding the items with negative discrimination, the subscales of the developed questionnaire have high internal consistency and reliability values that lead to the conclusion that the TAIA scale is of high quality.

The third goal was to examine the conformity of the theoretical model to collected data. Construct structure is supported by the data as there are no factors with insignificant variances. This means that all the suggested factors exist. However, the measures of model fit did not have satisfactory values. This indicates the insufficient model quality and structure of the relation between items and latent factors need to be reexamined. One possible reason for this is that some observed variables measure not only their latent variables. The results of modification indices analysis support this idea.

Additionally, it may be supposed that some factors initially proposed by the theoretical model can merge into a single latent variable. In particular, predictability and consistency, and faith and dependability may represent two sides of one attribute of trust. Although the correlation between latent variables does not significantly improve model quality, this hypothesis should be verified on a large sample.

TAIA total score slightly, but significantly positively correlated with GTS score that supports the hypothesis about the relation between GTS and TAIA. The GTS score correlated with TAIA subscales suggests that the developed subscales measure various aspects of trust. The correlational data supported the assumption that there are correlations between the experience of dealing with artificial intelligence and TAIA total score. The absence of correlation between the understanding subscale and data of experience dealing with AI technologies may be related to a low loading of the subscale in TAIA structure. These findings assume that despite the high consistency of this subscale, the factor of understanding how systems work does not influence the level of trust.

General discussion and further research

The result of the first iteration of the TAIA Scale showed satisfactory results. A new psychometric construct structure was elaborated based on classical theoretical concepts of trust and modern investigations in the digital trust research area.

The data collected support the theoretical structure of the construct of interest: there are no fake factors, and all regression coefficients on the CFA model are significant. Although based on the statistical properties, the quality of the current model is slightly lower than conventional thresholds, and it may be improved by correcting the item stems and further restructuring the model.

The quality of the construct of interest is supported by the statistical properties of the subscales and individual items since their values are greater than conventional ones. Furthermore, the consistency of understanding subscale is higher since it is a knowledge scale.

Evidence of the validity of the scale consists of the following elements. Conceptual validity arises from the fact that well-established trust research approaches were used to develop the structure of the construct. As the type of construct validity, the inner consistency was evaluated, and high values were obtained. During the quantitative approbation of the items, face validity was tested using probing techniques, and satisfactory results were also obtained. Correlation analysis was conducted to evaluate convergent and criterion validity and significant correlation of TAIA scores and GTS and data about the experience of dealing with intelligent technologies received from self-reported answers. Significant correlation coefficients were obtained that led to the conclusion of the scale's validity.

Further research will finalize the model structure to achieve satisfactory values for model fit. The item statements will be reviewed and revised to enhance the quality of measurement.

Conclusion

The ultimate goal of this study is to develop the TAIA Scale so that it can be successfully used for research on a Russian sample. Within the general purpose, the study focuses on achieving scale validity and on specifying and obtaining a correlation between the construct of interest and the GTS, and the quality of experience of dealing with artificial intelligent agents. The statistical methods from the Classical Test Theory are used to assess the scale quality. Cronbach's α is used to estimate reliability while measuring properties are examined through difficulty and discrimination coefficients

analysis. These measures showed the high quality of the developed tool. Pearson's coefficient is used to investigate correlations with related constructs, and a high significance of coefficient was obtained.

Since there were no other valid and reliable scales measuring the same or similar constructs, external validation was implemented based on the participants' experience of dealing with AI technologies. As a result, the scale positively correlates with self-reported data of such experiences and the GTS.

This study provides the first attempt to develop the TAIA Scale. The scale can be used for further psychological research. The major limitation of the study is the absence of agreement about the concept of trust. The small size of the sample may also affect the results. After reexamining the item stems, the study will be continued with a professionally curated sample to obtain more reliable results.

References

Akimova A. Yu., & Oboznov, A. A. (2016). Man's trust and distrust to technical equipment. *Psychological Journal, 6*, 56-69.

Allen, M.J., & Yen, W. M. (2002). Introduction to Measurement Theory. *Long Grove*. IL: Waveland Press.

Baier, A. (1986). Trust and Antitrust. *Ethics, 96(2)*, 231–60. Retrieved from https://www.jstor.org/stable/pdf/2381376.pdf

Bauer, Paul C., Conceptualizing Trust and Trustworthiness (November 5, 2019). Working paper published in: Political Concepts Working Paper Series, No. 61; Currently under review for "Trust Matters: Cross-Disciplinary Essays", Available at SSRN: https://ssrn.com/abstract=2325989 or http://dx.doi.org/10.2139/ssrn.2325989

Bellman, R. (1978). *An introduction to artificial intelligence: can computers think?* San Francisco: Boyd & Fraser Publishing Company; Thomson Course Technology.

Bisen, V. S. (2019, December 9) Where Is Artificial Intelligence Used: Areas Where AI Can Be Used. Retrieved from https://medium.com/vsinghbisen/where-is-artificial-intelligence-used-areas-where-ai-can-be-used-14ba8c092e73

Borum, R. (2010). The science of interpersonal trust. *Mental Health Law & Policy Faculty Publications*. 574. https://digitalcommons.usf.edu/mhlp_facpub/574

Castells, M. (2014). The impact of the Internet on society: a global perspective. *Change, 19*, 127-148.

Chen, J. Y., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science, 19(3)*, 259-282.

Faulkner, P., & Simpson, T. (Eds.). (2017). *The philosophy of trust*. Oxford: Oxford University Press.

Hardin, R. (2006). *Trust*. Cambridge: Polity.

Hauben, M. & Hauben R. (1997). The Vision of Interactive Computing And the Future. In M. Hauben & R. Hauben (Eds.), *Netizens: On the History and Impact of Usenet and the Internet*. Los Alamitos, CA: IEEE Computer Society.

Kramer, R. M., & Isen, A. M. (1994). Trust and distrust: Its psychological and social dimensions. *Motivation and Emotion, 18(2)*, 105-107.

Larsen, K. K. (2018, December 3) TRUST THOU AI? Retrieved from https://aistrategyblog.com/2018/12/03/trust-thou-ai/

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50-80.

Leont'ev, A. N. (1978). *Activity, Consciousness and Personality*. Englewood Cliffs, New Jersey: Prentice Hall.

Lyons, J. B., & Havig, P. R. (2014, June). Transparency in a human-machine context: approaches for fostering shared awareness/intent. In *International Conference on Virtual, Augmented and Mixed Reality* (pp. 181-190). Springer, Cham.

Nestik, T. A. (2018). Socio-Psychological Predictors of Person's Attitudes toward New Technologies. *Information society: education, science, culture and future technologies, 2*, 309-319.

Norman, D. (2013). *The Design of Everyday Things: Revised and Expanded Edition.* New York: Basic books.

Paxton, P., & Ressler, R. W. (2018). Trust and participation in associations. In E. M. Uslaner (Ed.), *The Oxford handbook of social and political trust*, (pp. 149-172). Oxford, UK: University Press.

Poole, D., Mackworth, A., & Goebel, R. (1998). *Computational Intelligence. A logical approach*. New York: Oxford University Press.

Pynadath, D. V., Barnes, M. J., Wang, N., & Chen, J. Y. (2018). Transparency communication for machine learning in human-automation interaction. In J. Zhou & F. Chen (Eds.), *Human and Machine Learning* (pp. 75-90). Springer, Cham.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rao A. & Cameron E. (2018, July 31). The future of artificial intelligence depends on trust. *TECH & INNOVATION. Strategy+business.* Retrieved from *https://www.strategy-business.com/article/The-Future-of-Artificial-Intelligence-Depends-on-Trust?gko=af118*

Revelle, W. (2020) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych

Rich, E. & Knight, K. (1991). *Artificial Intelligence.* New York: McGraw Hill.

Rosenberg, M. (1956). Misanthropy and political ideology. *American sociological review, 21(6)*, 690-695.

Rosenberg, M. J., Hovland, C. I., McGuire, W. J., Abelson, R. P., & Brehm, J. W. (1960). *Attitude organization and change: An analysis of consistency among attitude components. (Yales studies in attitude and communication.).* Yale Univer. Press.

Rosseel Y (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48(2),* 1–36. https://www.jstatsoft.org/v48/i02/.

Russell, S., & Norvig, P. (1995) *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice-Hall.

Salin, E. D., & Winston, P. H. (1992). Machine learning and artificial intelligence: an introduction. *Analytical chemistry (Washington, DC), 64(1),* 49A-60A.

Trust in Artificial Intelligence. Transform your business with confidence. (2018). *KPMG LLP* Retrieved from https://assets.kpmg/content/dam/kpmg/uk/pdf/2020/01/trust-in-artificial-intelligence-paper.pdf

Uslaner, E. M. (Ed.). (2018). *The Oxford handbook of social and political trust*. Oxford, UK: University Press.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Willis, G. B. (2005). *Cognitive interviewing. A tool for improving questionnaire design.* Thousand Oaks, CA: Sage Publications, Inc.

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and emotion, 18(2),* 129-166.

Zinchenko, V. P, & Panov, D. Yu. (1962). Uzlovye problemy inzhenernoi psikhologii [Key issues of industrial psychology]. *Voprosy Psikhologii [Issues of Psychology], 5,* 15-30.

Zinchenko, V. P, Leontiev, A. N., & Panov, D. Yu. (1964). Problemy inzhenernoi psikhologii [Issues of industrial psychology]. In A. N. Leontiev, V. P Zinchenko, & D. Yu. Panov (Eds.), *Inzhenernaya psikhologiya [Industrial psychology]* (pp. 5-23). Moscow: Moscow University Press.

Contact details

Anton Angelgardt

HSE University (Moscow, Russia). Laboratory for Cognitive Psychology of Digital Interfaces User. Research Assistant;

E-mail: aangelgardt@hse.ru

Elena S. Gorbunova

HSE University (Moscow, Russia). School of Psychology, Laboratory for Cognitive Psychology of Digital Interfaces User. Associate Professor, Laboratory Head;

E-mail: esgorbunova@hse.ru

Maria Chumakova

HSE University (Moscow, Russia). School of Psychology. Associate Professor;
E-mail: mchumakova@hse.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.