Transparency and Public Communication Foster Trust in AI Companies

Benjamin Prissé[1]*, Assel Mussagulova[2], Jun Quan Ho[1]

Abstract
This study examines how organizational characteristics of companies producing Artificial Intelligence (AI) technologies influence public trust through a vignette-based experimental design. Building on prior frameworks of trust, we focus on five sub-dimensions of trust: Benevolence, Standards and Guidelines, Data Quality, Reliability, and Transparency, each with three different levels. Results indicate that Transparency and Benevolence are the most significant drivers of trust. Organizations that provide clear explanations of their AI technologies and demonstrate societal accountability by seeking and incorporating public feedback are viewed more favorably. Adherence to external standards, such as national or international guidelines, further enhances trust, while technical performance and data quality are less influential, as participants assume the technology is functioning adequately for their limited use. We conclude that transparent practices, societal engagement, and institutional collaboration will foster public confidence in companies producing AI technologies.

JEL Classification: C9; O3

Keywords
trust in AI — dimensions of trust — AI companies — vignette experiment

[1] *Lee Kuan Yew Centre for Innovative Cities, Singapore University of Technology and Design.*
[2] *The School of Social and Political Sciences, University of Sydney.*
***Corresponding author**: benjamin.prisse@gmail.com.

Introduction

Trust in Artificial Intelligence (AI) is a topic that has been on the rise since AI-enhanced technologies became common in a variety of domains. Trust-based human-AI interactions are becoming increasingly common in public services (e.g., chatbots), healthcare, marketplaces (e.g., e-commerce), education, and financial services. In a bid to understand how human-machine tension affects trust and hence effective deployment of AI technologies various research strands have emerged. One prominent stream is on trust in technology ((Cabiddu, Moi, Patriotta, and Allen 2022); (Ghazizadeh, Peng, Lee, and Boyle 2012); (Lee and Moray 1994); (Lee and See 2004); (Zuboff 1988)) and innovation. The other perspective recognizes the importance of trust in the innovating firm and its communication (e.g., (Brock 1965); (Chiesa and Frattini 2011); (Ram and Sheth 1989); (Sternthal, Dholakia, and Leavitt 1978)). Much of the debate around the acceptance of new technologies has revolved around technology in general (Hooks, Davis, Agrawal, and Li 2022) and is not AI-specific. The research that does address specific AI technologies relies heavily on self-report measures which tend to be biased. Moreover, when it comes to trust in organizations that design and deploy AI, little research has been done about the specific dimensions of these organizations that affect trust in AI technologies. We propose to bring these two perspectives together by examining individual trust in AI and trust in organizations designing and deploying AI in a quasi-experimental setting.

An often-overlooked perspective on trust in AI is that these innovations are designed and deployed by organizations such as governments, universities, companies, etc. Hence whether people trust AI is likely to be influenced by whether people trust such organizations. Thus, one of the foci of this project is trust in organizations, specifically the impact of organizational reputation on trust. In doing so, we make a novel contribution to the field as there are currently no studies focusing specifically on the role of organizational reputation in enabling public trust in AI. As several case studies have shown, there is a reason for the public to be critical of such companies and public institutions because the consequences of how they use and implement AI and big data technology are not always clear and may, in fact, disadvantage certain socioeconomic groups (e.g. (Eubanks 2018); (O'neill 2016)). One aspect that affects individuals' judgment in this regard is that of organizational reputation.

A case in point is Facebook's controversial use of face recognition of their user base, based on AI, which would create a 'faceprint' of users who opted into the service so that all their images on the Facebook platform are automatically tagged. This technology was controversial because, at the time of implementation, it was unclear how this data would be used, and stored, and what the potential consequences of it would be. Reacting to public backlash, Facebook decided to put

the technology on hold because of "growing concerns about the use of this technology as a whole" and "many concerns about the place of facial recognition technology in society" as well as the fact that there were no clear guidelines and regulations in place. It is thus fair to say that not only what a company does (e.g. the technologies they roll out) but also how they do so (e.g. how they communicate about technology and the transparency they provide) is paramount for the reputation and legitimacy of a company (cf. (Anteby 2010)) and, consequently, how the public will trust AI.

Organizational reputation is used to refer to stable, shared affective evaluation and perception of the organization (Lievens 2017), in other words, how people feel about an organization. It is important because how various stakeholders feel about the organization, influences how they respond to the organization and the technologies it employs. Consequently, the reputation of an organization might have potentially favorable and unfavorable consequences. For instance, the reputation of an organization might serve as a signal of product quality, translate into attractiveness for investors, and potential employees, as well as improved organizational performance (Lievens 2017). Similarly, organizational reputation may impact human trust in AI. For instance, (Hengstler, Enkel, and Duelli 2016), in an analysis of nine case studies from the health and transportation industries concluded that firms promote trust in AI by connecting it to the reputation of the developing organization which could be related to the perceived moral standards of the algorithm thus affecting the trust of individuals. However, this study is limited to perceptions, without presenting how organizational reputation increases or decreases trust. (Frank and Otterbring 2023) investigated the link between consumer trust in a company and the intentions of consumers to adopt high (vs low) autonomy AI services and found a significant and positive relationship. Although real companies were used, the study did not capture the underlying mechanisms of trust and thus proved the gap in the current literature. On the other hand, research into trust in AI is developing into promising new avenues. (Bedué and Fritzsche 2022) expanded the (Mayer, Davis, and Schoorman 1995) framework of organizational trust which rests upon three constructs: ability, integrity, and benevolence, by identifying several specific sub dimensions of trust (see Figure 1). This framework captures important prerequisites of trust in AI which straddles organizational characteristics such as system and service quality, data quality, certifications, and ethical behavior, among others. Therefore, there is a whole body of work on organizational reputation and, likewise, a body of work on trust in AI. How these relate and interact, however, is yet unclear although paramount in understanding how public trust in AI technologies develops.

Methods

Our paper is based on (Bedué and Fritzsche 2022), who interviewed decision-makers in the industry with a solid understanding of AI to determine which dimensions are important for the general population to trust AI technologies. They identified three dimensions to be important to subjects: *Ability*, *Integrity* and *Benevolence*. The first refers to a company ability to perform an intended behavior and to signal this competency, the second refers to acting according to moral principles, the third refers to acting for the good of others. These dimensions are decomposed by the authors in additional sub-dimensions:

(i) *Ability* is further decomposed in *Access to knowledge*, which refer to the possibility that information technology provider share their knowledge with institutions and individuals to help the implementation of the technology in society; *Transparency*, which refer to how much the company developing the AI technology can understand how the technology makes its decisions and how it can be influenced; *Explainability*, which refers to the ability of the company to make the functioning of the technology understandable to the layperson; *System and service quality*, which refer to the ability of a company to control the training of the AI system to produce good results and deliver them to their customers; *Data Quality*, meaning the ability to train the AI technology on high-quality data that avoids perpetuating human and societal biases; and *Reliability*, which refer to the number of environments in which the technology can perform its task effectively.

(ii) *Integrity* is sub-decomposed in *Standard and Guidelines*, which refers to the ability of the company to adhere to self-imposed standards and guidelines that supplement official regulations and establish the competitiveness of the company; *Certifications*, meaning the ability to easily signal the quality of the AI technology to customers; *Government regulation*, meaning the possibility for a company to act according to existing laws and regulations.

(iii) *Benevolence* can be decomposed in *Social responsibility*, meaning the ability of the company to participate in a dialogue with others actors of society and listen to their feedback on how the technology should be developed; *Ethical behavior*, which refers to the ability to make morally responsible decisions in the context of actions that will significantly impact society; *Sustainability*, meaning the ability to develop technologies respecting the environment. Altogether, these dimensions determine the perceived benefits and risks individuals associate with using the technology. They form a theoretical framework for factors influencing trust in AI. The goal of this study is to investigate which of these dimensions are judged important by individuals for trusting the company producing AI technology. We will do so by using the methodology of vignette experiment for that. We will now explain how we built the vignettes.

Following the recommendations of (Aiman-Smith, Scullen, and Barr 2002) to avoid exceeding seven dimensions and thereby saturating the cognitive limits of individuals, we selected the most important dimensions suggested by (Bedué and Fritzsche 2022) that would also likely be understood by experimental subjects from the general population. We first selected *Benevolence* because it seemed evident that the first concern of a layperson would be the attitude of the company toward society. In this dimension, we selected *Social respon-*

sibility, as exchanges between companies and individuals is likely a salient concern to the layperson. We excluded *Ethical behavior*, considering it a necessary pre-requisite for any company behavior, and *Sustainability*, which experts consider currently too vague. We therefore define *Benevolence* according to *Social Responsibility* in our experimental design. The second dimension we selected is *Integrity*, since the layperson will likely ask for evidence of the morality of the behavior of the company. We selected the dimension *Standards and Guidelines* as a important ongoing debate in the industry and excluded both *Certifications* and *Government*, based on expert opinions that these sub-dimensions are currently lacking empirical meaning. We therefore define *Integrity* as *Standard and Guidelines* in our experimental design. Last, we selected the most important components of *Ability*, assuming that they would not necessarily be fully understood by the layperson, but that the opinion of the layperson regarding these topics is precisely the point of interest of vignettes experiment. We first selected *Data Quality* due to the importance of this topic in the performance of AI systems, and consequently excluded *System and service quality* because it was likely to overlap with *Data Quality* in the understanding of the layperson. We then selected *Reliability* because it encompasses the performance aspect of the technology, and it was essential to measure the opinions of subjects on this topic. Finally, we selected *Transparency* upon *Explainability* and *Access to knowledge* because we believed that what would matter to the layperson is whether a societal authority understands and validates the technology, not whether he understands the technology. In conclusion, we selected the sub-dimensions *Data Quality*, *Reliability* and *Transparency* within the *Ability* dimension.

Regarding the number of values, we decided to settle for the standard choice of three values for each dimension: positive, neutral, and negative. We therefore have a total of 3*3*3*3*3 = 243 possible vignettes. It is impossible for experimental subjects to evaluate that many vignettes; therefore, we determined how many treatments to use in separating the vignettes. Following (Atzmüller and Steiner 2010), we consider that individuals are likely to experience fatigue after approximately 40 scenarios. Following (Aiman-Smith, Scullen, and Barr 2002), we consider that participants need a training period of approximately 10 vignettes before providing accurate evaluations. We therefore decided to present subjects with 40 vignettes, with the first 13 vignettes being training vignettes that will not be included in the final analysis, and the next 27 vignettes being our vignettes of interest. The training vignettes were distinct from the vignettes of interest in each treatment, and subjects were not made aware of the training period so as not to influence their evaluations. Although subjects might experience fatigue by the end of the experimental design, we consider that having too many vignettes is better than having too few, following the recommendations of (Cooksey 1996) and (Stewart 1988), who suggest maintaining at least a 5:1 ratio between the number of vignettes in the experiment and the number of dimensions per vignette. We therefore allocated the vignettes into 9 different treatments, with each vignette assigned to a single treatment. To ensure that subjects see a representative sample of vignettes, the subsamples of vignettes in each treatment are pseudo-randomly created. We coded a program in the RStudio statistical software to allocate the 243 vignettes to the 9 different treatments, meaning that each vignette is present in a single treatment. This program is available on Mendeley Data, along with the dataset, statistical analyses, and supplementary materials[1]. This number of treatments was the minimum required for the algorithm to allocate all possible combinations of dimension values across treatments, ensuring that each dimension value was represented at least twice during the training period and six times during the evaluation period.[2]. Finally, we decided not to build a story to describe the company, in order to minimize the effort that subjects have to make. We only presented them with the different values of the dimensions.

We will now present the different values of dimensions and explain their justification:

Benevolence: We considered that AI companies might not necessarily understand the broader societal consequences of their technology, as their activities are primarily driven by monetary benefits. While they might foresee the most immediate and predictable outcomes, they are likely to overlook the long-term and indirect effects. We also considered that companies might not interact directly with the general population, given that their motivations and purposes may not be easily understood by the laypeople. This lack of understanding could lead to resistance from groups attempting to block the development of their technologies. Technologies like ChatGPT and Midjourney were imposed on the population, significantly impacting the education system and the creative middle-class job market (writers, artists). These sectors were forced to adapt to the existence of these technologies, rather than being allowed to gradually assimilate them. As AI technologies become more accepted in society, it will be necessary to understand how people experience the technology during its release. We therefore consider that the *Benevolence* dimension is about understanding the consequences of the technology on society and being receptive to the consumer feedback. We judge the first factor more important than the second because it is always difficult to conceive the macro-consequences rather than the micro-consequences of human actions. We consider that the positive value of the dimension is the company understanding the consequences of its technology and updating its development by listening and processing the feedback of customers. We consider that the neutral value of the dimension is the company understanding the consequences of the technology but not considering the feedback of customers. We consider that the negative value of the dimension is the company not

[1] Reserved DOI: 10.17632/8rg46xfw4m.4

[2] The two codes do not function together, meaning that we needed to verify that vignettes in the training period are not also present in the evaluation period, and manually correct the samples.

understanding the consequences of the technology on society and not considering the feedback of customers.

- (Positive) The company envisions the benefits of its technology for society and takes feedback from the customer into account to guide its development.
- (Neutral) The company envisions the benefits of its technology for society and does not take feedback from the customer into account to guide its development.
- (Negative) The company does not envision the benefits of its technology for society and does not take feedback from the customer into account to guide its development.

Standard and Guidelines: The topic is currently debated due to the lack of standards and guidelines for AI technologies, and the difficulty of establishing them for specific AI applications. Using general standards and guidelines that imitate those currently used in the industry might be sufficient, as these technologies inherently perform general tasks and simply optimize their execution. We therefore consider that the starting point for *Standard and Guidelines* is their existence. Since establishing these standards and guidelines will require the participation of external actors, such as governmental institutions or international regulators, we consider that the second criterion for this dimension is the participation of said actors and at which scale these dimensions are implemented. The positive value of this dimension is, therefore, the existence of standards and guidelines established with the participation of external actors from the industry at international level. The neutral value of this dimension is the existence of standards and guidelines with the participation of external actors at the national level. The negative value of this dimension is the absence of standards and guidelines defined by external actors, and therefore the commitment of the company to an internal code of conduct.

- (Positive) The company adheres to international industry-specific standards and guidelines.
- (Neutral) The company adheres to national industry-specific standards and guidelines.
- (Negative) The company adheres to internal standards and guidelines.

Data Quality: The topic of data quality is also currently debated. The accuracy of AI technologies is intrinsically linked to the quantity and quality of data used to train them. Although the general understanding is that the quantity of data is the main criterion, as it naturally helps the AI converge to the generally optimal answer, the quality of data remains an important factor because poor data quality can perpetuate human or societal biases through the results of the AI technology. For example, AI technologies have been demonstrated to discriminate against Black people in America. However, building a qualitative dataset is costly, as it requires human expertise to determine which data are important to collect and how to use them, and possibly the creation of new data collection methods. Such tasks might be more easily accomplished by companies than by public institutions because the former will have more liberty and incentives to do so. We therefore consider that the first criterion in the *Data Quality* dimension is quality because it is the hardest to satisfy. The second criterion is the quantity of data, as accumulating a large quantity of data is the easiest task. The positive value of this dimension is, therefore, the AI technology being trained on a qualitative dataset created within the company. The neutral value is the AI technology being trained on a large quantity of data obtained through the publicly available database of an institution. The negative value is the AI technology being trained on a limited amount of data obtained from a publicly available database. We omit the case of an AI technology trained on both a qualitative and quantitative dataset, considering that the measurement of interest is whether individuals value more the quantity or quality of data, and how subjectively important they believe these parameters to be compared to the baseline.

- (Positive) The company trains its technology on a limited amount of data it has designed specifically for this task.
- (Neutral) The company trains its technology on a large amount of data collected from publicly available sources.
- (Negative) The company trains its technology on a limited amount of data collected from publicly available sources.

Reliability: This dimension refers to the performance of AI technology. Traditional definitions of performance are not suitable for AI, as it surpasses human abilities in both the quality of results and the time taken to achieve them. Mechanical failure is not considered a component of AI performance, as the algorithm focuses on interpreting the data it receives and responding accordingly. Instead, AI performance is better defined by the number of different environments it can handle. According to its intended use, an AI technology may face various logical requests, emotional reactions, or natural environments, and must respond correctly in each circumstance, despite limited or partially interpreted programming data that might not cover every situation. The inability of the technology to respond accurately, or worse, the possibility of responding erroneously, could potentially endanger its users. Thus, a measure of AI performance is the number of environments it can manage. A reliable AI should perform well in the standard environments relevant to its intended use and should also have the capacity to handle less frequent and potentially more complex situations that might arise in these environments. The *Reliability* dimension is exclusively concerned

with this factor, considering that a top-performing AI technology would handle most complex environments it is potentially confronted with, while a bottom-performing AI technology would manage regular environments but struggle with any additional and unexpected difficulty. The positive value of this dimension is therefore an AI technology who can handle the complexity of many environments. The neutral value is an AI technology who can handle the complexity of some environments. The negative value is an AI technology who can only handle standard environments. We assume that this scale of evaluation is subjective, relative to the average degree of difficulty that the type of AI technology would handle.

- (Positive) The company designs technology that can perform its tasks in standard and a variety of complex environments.
- (Neutral) The company designs technology that can perform its tasks in standard and certain complex environments.
- (Negative) The company designs technology that can perform its tasks in standard environments.

Transparency: This dimension refers to the amount of explanations given by the company about its technology. The starting point of providing explanations is whether the company can convey these explanations efficiently to the interlocutor, but since the question of explaining AI technologies to the audience is an ongoing research field, we will not treat this aspect in our vignettes and will assume that companies are able to communicate efficiently to consumers. Because AI technologies are complex intellectual objects, either governmental agencies or international regulators will have to review these technologies to verify that they pose no particular harm to society. Such functioning implies "open source" access to these technologies, meaning the full disclosure of how the AI technology functions, but this might not be particularly useful to the layperson except for understanding that the technology has been thoroughly verified by qualified experts. However, the competitive constraints of the industry are such that it is unlikely that companies will voluntarily disclose the functioning of their technology, since their rivals would likely use their work for their own profit. Current technology like ChatGPT does not provide any information on how it functions. A neutral step would be to provide enough information to customers and institutional actors to explain how the model behind the AI technology works, while keeping the data or particular technical aspects of the model confidential to protect the know-how of the company. We therefore consider that the *Transparency* dimension is exclusively composed of the amount of information that the company is ready to share. The positive value is a full disclosure of the model behind the AI technology. The neutral value is a partial disclosure of the model behind the AI technology. The negative value is not disclosing anything about the model behind the AI technology. The last case corresponds to the current situation in the AI industry, therefore allowing this dimension to measure how much individuals value this information.

- (Positive) The company provides full explanations on how its AI technology work.
- (Neutral) The company provides selective explanations on how its AI technology works.
- (Negative) The company provides no explanations on how its AI technology works.

The experimental design includes a second part where subjects are presented with two vignettes simultaneously, 20 times. In these vignettes, all dimensions are set to neutral values except for one different dimension in each vignette, which is either positive or negative. Subjects are then asked to indicate which of the two companies they trust more. The comparisons are the same in all treatments, with the only change being the order in which these comparisons are presented to control for order effect. The objective of this part of the experiment is to evaluate which positive and negative values of the dimensions subjects consider the most important by making direct comparisons between them.

Once subjects completed the second part of the experiment, they participated in two economic tasks: the Dictator Game and the Ultimatum Game. In both games, Player A was endowed with 10 tokens, while Player B had 0 tokens. Tokens were convertible into money at a given exchange rate at the end of the experiment. However, these tasks were not separately incentivized, as participants received a fixed payment for completing the vignette experiment, and these tasks were secondary to our study. This is unlikely to affect the results, as (Larney, Rotella, and Barclay 2019) show that stake size in these games has almost no effect on outcomes for standard monetary amounts. Experimental subjects always took the role of Player A, whereas Player B was a fictional AI system. In the Dictator Game, Player A chooses how many tokens to send to Player B, and the game ends. This game measures participants' altruism. In the Ultimatum Game, Player A proposes a division of the tokens, and Player B can accept or reject the offer. If Player B accepts, tokens are allocated according to Player A's proposal; if Player B rejects, both players receive nothing. These games respectively measure altruism and strategic thinking, which we control for in our regressions. Interested readers can consult (Engel 2011) for a meta-analysis on the Dictator Game and (Camerer 2003) for a review of the Ultimatum Game. After completing the games, subjects answered a questionnaire collecting their sociodemographic information (e.g., age, gender, occupation).

The second questionnaire is an AI questionnaire specifically designed for the experiment. This questionnaire is subdivided into five items. The first item presents subjects with a list of ten characteristics and asks them whether they believe AI possesses these characteristics. The second item presents subjects with a list of ten technologies and asks whether they believe these technologies utilize AI. The third item presents

subjects with ten technologies that use Artificial Intelligence and asks them to indicate their current development status. The fourth item presents subjects with the same ten technologies as the third item and asks them to assess the anticipated impact of these technologies on their employability. The fifth item presents subjects with the same ten technologies as the third item and asks them to assess their trust in these technologies. Overall, the goal of this questionnaire is to assess whether subjects have a comprehensive understanding of AI, whether they are knowledgeable about the current use and developments of AI, and whether they hold positive or negative expectations regarding AI. We use these responses as control variables in our regressions to account for participants' knowledge of and beliefs about AI.

The experiment was programmed with oTree and conducted online. Participants were recruited online through invitations posted in Telegram groups dedicated to the recruitment of experimental participants in Singapore. Individuals who responded were registered for an experimental session and then received a link to a Zoom session. We recruited 702 participants from Telegram groups and collected data between February 1 and February 29, 2024, until reaching the targeted sample of 600 individuals who joined our experimental sessions and completed the experiment.

Once subjects connected to the experimental session, they were sent an individual link to the experiment in the Zoom chat. They opened the link in their browser and provided their answers to the experiment. After completing the experiment, they provided their personal information and were sent a €20 voucher via email as payment. Subjects confirmed receipt of the voucher and exited the experiment. We will now turn to the analysis of results.

We estimated the following regression model:

$$y_i = \beta_0 + \beta_1 * X_{id} + \beta_2 * X_{ie} + \beta_3 * X_{ia} + \beta_4 * X_{is} + \varepsilon_i \quad (1)$$

Where y_i is the grade attributed to a vignette by an individual; X_{id} is the estimated impact of each dimension compared to the baseline negative value; X_{ie} is the vector of control variables for economic tasks; X_{ia} is the vector of control for AI questionnaire, X_{is} is the vector of control for sociodemographics characteristics and ε_i is the error term.

The control variables for economic tasks X_{ie} are the amount of money sent in the Dictator Game ($SentAmountDG$) and the amount of money proposed in the Ultimatum Game ($SentAmountUG$). They respectively control for the altruism and strategic thinking of participants. These variables may influence the results, as altruistic individuals could assign higher grades to vignettes due to their benevolent disposition. Additionally, strategic decision-making may lead participants to send positive amounts to others to please them, mirroring the potential inclination to give higher grades to vignettes in response to a perceived demand by the experimenter.

The control variables for the AI questionnaires X_{ia} are the number of correct answers (from 0 to 10) in the first part of the AI questionnaire ($ScoreAI1$), the number of correct answers (from 0 to 10) in the second part of the AI questionnaire ($ScoreAI2$), the number of correct answers (from 0 to 10) in the third part of the AI questionnaire $ScoreAI3$, the average likeliness (from 1 to 7) of seeing their employability being influenced by different AI technologies ($ScoreAI4$) and their average trust (from 1 to 7) on different AI technologies ($ScoreAI5$).

The sociodemographic controls X_{is} are separated into several subcategories. The first subcategory comprises all the traditional controls that are Age, $Male$ and nationality of subjects, with baseline level being a Singapore citizen by birth. The second subcategory is the level of education of the subject, the third subcategory is the level of education of his father, and the fourth category is the level of education of his mother, all with the baseline level being a Polytechnic degree. The fifth subcategory is made of others sociodemographics variables with a potential impact on results, such as being the $OnlyChild$ in the family, the field of study with baseline level being Business and Administration, or being a student of the $SUTD$ university in which the research was conducted.

Sociodemographics

Table 1 presents a summary of the sociodemographic characteristics of the sample. The average age of participants is 32.10 years, 43.5% are male, and 16.7% are the only children of their family. Regarding their nationality, 81.5% of participants are Singaporean by birth, and 8.3% are Permanent Residents of Singapore. Regarding the education of subjects and their families, 66.2% of subjects have a tertiary level of education, while 47.0% of their fathers and 43.9% of their mothers have attained the same level of education. Regarding their field of study, 27.8% of subjects studied Business and Administration, 17.0% studied Humanities and Social Sciences, and 18.2% studied Science and Related Technologies. Regarding their professional situation, 27.5% of subjects are unemployed, while 60.2% are employed. This unemployment rate is much higher than the typical 2% or 3% in Singapore, but it most likely suggests that most of these subjects are students. Regarding their wage, 33% of subjects earn less than $1,000 per month, while 42.5% earn between $1,000 and $5,000 per month. Regarding their religious beliefs, 32.7% of subjects are Buddhists, 20.3% are non-Catholic Christians, and 31.2% have no religion. Regarding their economic preferences, subjects send an average of 3.333 points in the Dictator Game and 3.886 points in the Ultimatum Game. Regarding their knowledge of AI, subjects correctly answer an average of 8.337 questions on the first AI questionnaire, 7.807 questions on the second AI questionnaire, and 1.493 questions on the third AI questionnaire. Finally, their average score on belief in the impact on employability is 3.861, while their average score on trust in the application of technology is 4.544. We therefore conclude that the sample is slightly imbalanced toward male subjects and likely contains a higher proportion of students than the general population, but is otherwise a

Table 1. Tobit estimations of the impact of levels of dimension on the grade given to vignettes, with controls.

Variable	N	Mean	Std.Dev	Min	Max
Age	600	32.10	11.961	18	84
Gender	600	0.435	0.496	0	1
OnlyChild	600	0.167	0.373	0	1
Nationality (Singapore)	600	0.815	0.389	0	1
Nationality (Permanent Resident)	600	0.083	0.277	0	1
Education (Tertiary)	600	0.662	0.474	0	1
FatherEducation (Tertiary)	600	0.328	0.470	0	1
MotherEducation (Tertiary)	600	0.26	0.439	0	1
FieldofStudy (Business and Administration)	600	0.278	0.449	0	1
FieldofStudy (Humanities and Social Science)	600	0.17	0.376	0	1
FieldofStudy (Science and Related Technologies)	600	0.182	0.386	0	1
CurrentJob (Unemployed)	600	0.275	0.447	0	1
CurrentJob (Employed)	600	0.602	0.490	0	1
Income (Below $1000)	600	0.33	0.471	0	1
Income ($1000-$5000)	600	0.425	0.495	0	1
Religion (Buddhism)	600	0.327	0.469	0	1
Religion (Christian (Others))	600	0.203	0.403	0	1
Religion (No Religion)	600	0.312	0.464	0	1
SentAmountDG	600	3.333	2.946	0	10
SentAmountUG	600	3.886	2.427	0	10
*ScoreAI*1	600	8.337	1.318	2	10
*ScoreAI*2	600	7.807	1.228	2	10
*ScoreAI*3	600	1.493	1.122	0	5
*ScoreAI*4	600	3.861	1.573	1	7
*ScoreAI*5	600	4.544	0.986	1	7

diverse and representative pool of subjects.

Stage 1

Table 2 displays the results of the regressions. Column (1) shows the regression without controls, column (2) shows the regression with economic controls, column (3) shows the regression with AI questionnaire controls, column (4) shows the regression with sociodemographic controls, column (5) shows the regression with full controls. Table A1 displays the regressions with selected controls. We observe that the estimates of the impact of each level across dimensions are consistent across regressions, and we focus on column (1) for the interpretation of coefficients. We see that the average grade given to vignettes increases by 1.524 ($p < 0.001$) for the positive dimension of *Benevolence* compared to baseline, and by 0.689 ($p < 0.001$) for the neutral dimension of *Benevolence* compared to baseline. It suggests that individuals value companies who understand the benefit to society of their technology and give even further value to companies taking the feedback of customers into account. Regarding *Standard and Guidelines*, column (1) shows that the evaluation of companies increases by 0.619 ($p < 0.001$) with international standard and guidelines and by 0.473 ($p < 0.001$) with national standard and guidelines compared to the baseline of internal standards and guidelines, suggesting that individuals value the implementation of official standard and guidelines, but make little difference regarding the scale of implementation of these guidelines. Additionally, we see that subjects do not differentiate between low-quality data (limited amount collected from public sources) and high-quality data (limited amount collected by the company) since the positive dimension is not significant. We also see that the neutral value of this dimension, i.e. collecting a large amount of data from public sources, increases the evaluation of the company by 0.430 ($p < 0.001$). Now, we observe that *Reliability* is not important to the subjects. A high-level of performance increases the evaluation of the company by 0.140 with 1% significance ($p < 0.001$) and a neutral level of performance increases the evaluation of the company by 0.123 with 1% significance ($p < 0.001$). Finally, we see that *Transparency* matters to subjects. A positive value of the dimension increases the evaluation of the company by 1.817 ($p < 0.001$) compared to baseline, and a neutral value of the dimension increases the evaluation of the company by 0.873 ($p < 0.001$) compared to baseline. Subjects therefore give significant value to the company providing explanations about its technology, even if they are partial. Because the text does not suggest that these explanations are comprehensible to subjects, it therefore seems that simply explaining the technology publicly gives confidence to subjects. Additionally, we can quickly examine the control variables. Among interesting results, we find that *SentAmountDG* ($p = 0.025$ in (2) and $p = 0.050$ (5)) and *SentAmountUG* ($p = 0.037$ in (2) and $p = 0.047$ (5)) increase the grade given to companies. This suggests that higher altruism and strategic thinking in AI increase trust in the companies building AI. We also find that *ScoreAI*1 ($p = 0.025$ in (3) and $p = 0.006$ in (5)) and *ScoreAI*5 ($p < 0.001$ in both (3) and (5)) increase the grade given to companies, suggesting that altruism and strategic thinking are positively correlated with grades. This indicates that better knowledge of AI and greater trust in its real-life applications enhance trust in companies producing AI technologies. Finally, we find that older subjects give lower grades to companies producing AI technologies ($p = 0.007$ in (5)), suggesting that individual characteristics might also have an influence.

We conclude that *Benevolence* and *Transparency* are the most important dimensions to subjects. Having a vision of the benefits of technology for society, providing public information about the technology and taking feedback of individuals into account to guide the development of the technology are judged by individuals as the most important parameters to elicit their trust. We also see that subjects value the implementation of standard and guidelines by institutions, whether at local or international level. Interestingly, results regarding the data quality reveals that subjects do not understand this topic, outside of the intuitive understanding that more data is preferred to less. We also see that the performance of the technology does not matter, most likely because individuals assume that they will use AI technology for their standard needs. Finally, we see that a positive inclination toward AI or certain individual characteristics increases the level of trust in companies producing AI technologies.

We now include interaction terms in our regressions to account for the potential simultaneous effects between levels of dimensions. Table 3 displays the results of these regressions with a selected set of significant interaction variables. We see that the coefficient of *StandardNeu* now becomes 1% in all specifications, while it was previously 0.1% significant in all of them. We also see that the coefficient *DataNeu* loses signif-

Table 2. Tobit estimations of the impact of levels of dimension on the grade given to vignettes, with controls.

	(1) Grade	(2) Grade	(3) Grade	(4) Grade	(5) Grade
BenevolencePos	1.524***	1.531***	1.524***	1.508***	1.512***
	(0.074)	(0.073)	(0.073)	(0.071)	(0.070)
BenevolenceNeu	0.689***	0.695***	0.690***	0.677***	0.677***
	(0.053)	(0.053)	(0.053)	(0.050)	(0.050)
StandardPos	0.619***	0.613***	0.620***	0.624***	0.623***
	(0.050)	(0.050)	(0.050)	(0.050)	(0.050)
StandardNeu	0.473***	0.465***	0.466***	0.476***	0.464***
	(0.046)	(0.046)	(0.046)	(0.046)	(0.045)
DataPos	0.021	0.018	0.018	0.022	0.015
	(0.038)	(0.037)	(0.037)	(0.037)	(0.036)
DataNeu	0.430***	0.429***	0.427***	0.430***	0.427***
	(0.044)	(0.044)	(0.043)	(0.043)	(0.043)
ReliabilityPos	0.140***	0.140***	0.132**	0.149***	0.146***
	(0.044)	(0.043)	(0.043)	(0.041)	(0.040)
ReliabilityNeu	0.123***	0.127***	0.122***	0.124***	0.129***
	(0.037)	(0.037)	(0.037)	(0.035)	(0.035)
TransparencyPos	1.817***	1.809***	1.813***	1.819***	1.814***
	(0.069)	(0.069)	(0.069)	(0.068)	(0.068)
TransparencyNeu	0.873***	0.869***	0.873***	0.873***	0.872***
	(0.046)	(0.046)	(0.046)	(0.044)	(0.044)
Intercept	2.865***	2.433***	0.348	2.631***	-0.301
	(0.101)	(0.160)	(0.732)	(0.520)	(0.820)
Observations	16200	16200	16200	16200	16200
Number of Subjects	600	600	600	600	600
Estimation Method	Tobit	Tobit	Tobit	Tobit	Tobit
Log Pseudolikelihood	-34867.05	-34735.46	-34575.17	-34304.41	-33836.75
Economic Controls	No	Yes	No	No	Yes
AI Controls	No	No	Yes	No	Yes
Sociodemographics Controls	No	No	No	Yes	Yes

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

icance in all specifications except (2). These results therefore indicate that subjects do not attribute value to the quality of data, with a potential interpretation being that they do not understand the importance of the topic. Additionally, we see that *ReliabilityPos* loses significance in all regressions except the (5), indicating that subjects do not value an high level of performance from the technology. Regarding the interaction terms, we see that $BenevolencePos * DataNeu$ is positive and significant at 5% in columns (1) ($p = 0.014$), (2) ($p = 0.026$) and (3) ($p = 0.011$) and positive and significant at 1% in columns (4) ($p = 0.006$) and (5) ($p = 0.004$). It therefore means that participants positively value that the company has a vision of the benefits of its technology for society and take into account the feedback of customers when the technology is trained on a large quantity of public data. We also see that $BenevolencePos * TransparencyPos$ is positive and significant at 0.1% in all regressions. It therefore means that participants positively value the company having a vision of the benefits of its technology for society and taking into account the feedback of customers when the company releases full explanations on its technology. We have the same interpretation for these two interaction coefficients: subjects trust the official discourse of the company when it is supported by the actions of the company. We now see that $StandardPos * DataPos$ is positive and 1% significant in columns (2) ($p = 0.006$), (4) ($p = 0.008$) and (5) ($p = 0.008$), and positive and significant at 5% in columns (1) ($p = 0.012$) and (3) ($p = 0.016$), meaning that participants positively value companies training the technology on a limited amount of data specifically designed for this task when they operate within international standard and guidelines. Then, we see that $StandardPos * DataNeu$ is significant at 0.1% in all regressions, meaning that participants positively value a company training its technology on a large dataset collected from public sources when international standards and guidelines are implemented. Additionally, we see that $StandardNeu * DataNeu$ is positive and significant at 1% in columns (1) ($p = 0.003$), (2) ($p = 0.004$) and (3) ($p = 0.004$) and positive and significant at 0.1% in (4) and (5), meaning that participants positively value a company training its data on large amounts of public sources when the company operates within national standards and guidelines. These results suggest that subjects are sensible to particular combinations of data collection and institutional frameworks in giving their trust. Then, $DataPos * ReliabilityNeu$ is negative and significant at 1% in regressions (1) ($p = 0.005$), (2) ($p = 0.003$), (3) ($p = 0.002$) and (5) ($p = 0.010$), and negative and significant at 5% in (4) ($p = 0.028$). It therefore means that participants have a positive opinion of the company training its technology on a limited amount of data it has designed specifically for this task if the technology can perform its tasks in standard and certain complex environments. A potential interpretation is that subjects find it coherent to train the technology in specific environments and expect it to perform its tasks only in such environments. We also see that $DataNeu * TransparencyNeu$ is positive and significant at 5% in all columns ($p = 0.037$ in (1), $p = 0.022$ in (2), $p = 0.045$ in (3), $p = 0.021$ in (4), $p = 0.020$ in (5)). It means that individuals trust a company that trains its technology on a large amount of data collected from publicly available sources when the company provides selective explanations of how its technology functions. A potential interpretation is that having data and providing explanations are both factors of trust that mutually reinforce each other. Finally, $ReliabilityNeu * TransparencyNeu$ is negative and significant at 1% in regression (1) ($p = 0.009$), and negative and significant at 5% in regressions (2) ($p = 0.013$), (3) ($p = 0.012$), (4) ($p = 0.007$) and (5) ($p = 0.011$). It therefore means that participants have a negative opinion of the company training its technology on a limited amount of data it has designed specifically for this task if the company provides selective explanations on how its technology works. A potential interpretation is that subjects perceive a partial disclosure of information with targeted objective as possibly hiding something.

Overall, we see that subjects positively value coherence between the posture of the company and its actions, as well as the training of the technology and its use. They also value the company collecting its data inside an established institutional framework. They negatively value any perceived potential hidden agenda in the technology.

Table 3. Tobit estimations of the impact of levels of dimension on the grade given to vignettes, with selected interactions.

	(1) Grade	(2) Grade	(3) Grade	(4) Grade	(5) Grade
BenevolencePos	1.178***	1.186***	1.173***	1.153***	1.153***
	(0.121)	(0.120)	(0.118)	(0.118)	(0.114)
BenevolenceNeu	0.658***	0.641***	0.670***	0.666***	0.682***
	(0.116)	(0.115)	(0.116)	(0.110)	(0.108)
StandardPos	0.414***	0.414***	0.428***	0.390***	0.417***
	(0.116)	(0.113)	(0.115)	(0.112)	(0.111)
StandardNeu	0.340**	0.343**	0.350**	0.318**	0.348**
	(0.124)	(0.123)	(0.121)	(0.119)	(0.116)
DataPos	0.156	0.133	0.153	0.154	0.155
	(0.131)	(0.128)	(0.130)	(0.121)	(0.119)
DataNeu	0.265	0.286*	0.273	0.206	0.240
	(0.146)	(0.145)	(0.145)	(0.137)	(0.135)
ReliabilityPos	0.273	0.266	0.267	0.272	0.281*
	(0.158)	(0.154)	(0.157)	(0.144)	(0.139)
ReliabilityNeu	0.442***	0.440***	0.448***	0.411***	0.428***
	(0.120)	(0.118)	(0.119)	(0.112)	(0.111)
TransparencyPos	1.773***	1.746***	1.765***	1.777***	1.759***
	(0.139)	(0.137)	(0.139)	(0.132)	(0.131)
TransparencyNeu	1.043***	1.016***	1.041***	1.109***	1.100***
	(0.131)	(0.130)	(0.132)	(0.124)	(0.122)
*BenevolencePos * DataNeu*	0.229*	0.206*	0.232*	0.242**	0.245**
	(0.094)	(0.093)	(0.091)	(0.088)	(0.085)
*BenevolencePos * TransparencyPos*	0.396***	0.415***	0.400***	0.371***	0.382***
	(0.096)	(0.095)	(0.095)	(0.093)	(0.091)
*StandardPos * DataPos*	0.270*	0.292**	0.267*	0.268**	0.263**
	(0.108)	(0.106)	(0.110)	(0.101)	(0.0992)
*StandardPos * DataNeu*	0.406***	0.398***	0.384***	0.465***	0.429***
	(0.113)	(0.111)	(0.112)	(0.108)	(0.105)
*StandardNeu * DataNeu*	0.299**	0.287**	0.291**	0.345***	0.317***
	(0.101)	(0.099)	(0.100)	(0.096)	(0.095)
*DataPos * ReliabilityNeu*	-0.292**	-0.301**	-0.308**	-0.210*	-0.241**
	(0.103)	(0.101)	(0.102)	(0.096)	(0.093)
*DataNeu * TransparencyNeu*	-0.211*	-0.229*	-0.202*	-0.227*	-0.226*
	(0.101)	(0.100)	(0.100)	(0.099)	(0.097)
*ReliabilityNeu * TransparencyNeu*	-0.213**	-0.198*	-0.202*	-0.211**	-0.195*
	(0.081)	(0.080)	(0.081)	(0.078)	(0.077)
Constant	2.894***	2.472***	0.371	2.662***	-0.285
	(0.160)	(0.206)	(0.745)	(0.537)	(0.833)
Observations	16200	16200	16200	16200	16200
Number of Subjects	600	600	600	600	600
Estimation Method	Tobit	Tobit	Tobit	Tobit	Tobit
Log Pseudolikelihood	-34818.23	-34687.03	-34526.29	-34247.88	-33829.03
Economic Controls	No	Yes	No	No	Yes
AI Controls	No	No	Yes	No	Yes
Sociodemographics Controls	No	No	No	Yes	Yes

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Stage 1 - Training Periods

Table 4 displays the results of the regressions when considering only the training period. We see identical results, except for three variables: *DataPos* is now negative and significant at 0.1% in all regressions, therefore indicating that subjects initially associate a low quantity of data to something negative, but they rapidly understand that it is not necessarily the case, although they probably do not associate it with quality. *ReliabilityPos* loses significance and *ReliabilityNeu* becomes significant at 5% in all regressions. It therefore indicates that subjects initially do not give importance to the performance of the technology, but slightly differentiate between level of performance afterwards. Coefficients also indicate that subjects are initially less tolerant to the company not taking into account feedback of customers despite having a vision of the benefits of the technology for society (*BenevolenceNeu*), and initially give less value to international standard and guidelines (*StandardPos*).

Stage 2

We now analyze the results of Stage 2 of the experiment. Table 5 displays the results when comparing the positive levels of each dimension. We see that *Transparency* is preferred in 72.21% of the cases and that *Benevolence* is preferred in 64.88% of the cases, with positive pairings against *Standard*, *Data*, and *Reliability* in both cases. We also see that subjects prefer *Transparency* to *Benevolence* in the direct comparison between the two. These results suggest that the most important criterion for individuals to trust a company is that the company provides full explanations of how its technology works, closely followed by the company having a vision of the benefits of its technology for society and taking customer feedback into account. These results are consistent with the estimated coefficients in Stage 1 of the experiment, providing further evidence of the same findings. We additionally see that participants are less concerned with the performance of the technology and the quality of data. Regarding the performance of the technology, they choose *Reliability* 37.92% of the time on average, and only 24.39% of the time when excluding the comparison with *Data*. There are two potential and non-exclusive explanations for this phenomenon. The first is that subjects do not understand why they would use advanced capabilities of AI technology in their daily lives and therefore do not attribute value to the performance of the technology. The second is that subjects intuitively understand that they will not grasp the technicalities of AI technology and therefore prefer societal and institutional regulation to ensure that the technology will be beneficial to them. Regarding data quality, they choose *Data* 16.96% of the time on average. An interpretation of this phenomenon is that participants may not understand that this description fits the characteristics of qualitative data. They might even react negatively to learning that the company builds its own data rather than using public sources. This suggests that participants lack advanced knowledge of AI-related concerns, and that this ignorance might lead to issues when communicating with them because of their erroneous perceptions of what could be a problem.

Table 6 displays the results when comparing the negative levels of each dimension. We see that *Standard* and *Transparency* are the least selected dimensions with 39.96% and 37.42% of picks, respectively, suggesting that individuals have a negative opinion of companies that keep their operations internal and do not act within a collectively agreed-upon moral framework. This interpretation is further supported by *Benevolence* being the third pick, with 50.83% of selections, indicating that the absence of communication with the outside is negatively perceived, although not as negatively as the previous two dimensions. This may be because subjects find it justifiable not to continuously justify oneself to the layman. The most chosen dimension is *Reliability*, selected 64.13% of the time. This suggests that subjects do not view

Table 4. Tobit estimations of the impact of levels of dimension on the grade given to vignettes, with controls.

	(1) Grade	(2) Grade	(3) Grade	(4) Grade	(5) Grade
BenevolencePos	1.648***	1.649***	1.643***	1.655***	1.646***
	(0.077)	(0.077)	(0.077)	(0.076)	(0.075)
BenevolenceNeu	0.486***	0.486***	0.483***	0.498***	0.492***
	(0.061)	(0.061)	(0.061)	(0.061)	(0.061)
StandardPos	0.396***	0.388***	0.399***	0.429***	0.425***
	(0.068)	(0.068)	(0.067)	(0.067)	(0.065)
StandardNeu	0.522***	0.511***	0.525***	0.531***	0.528***
	(0.067)	(0.067)	(0.067)	(0.066)	(0.065)
DataPos	-0.224***	-0.220***	-0.217***	-0.210***	-0.199***
	(0.060)	(0.060)	(0.059)	(0.057)	(0.055)
DataNeu	0.291***	0.294***	0.294***	0.324***	0.326***
	(0.067)	(0.067)	(0.066)	(0.066)	(0.065)
ReliabilityPos	-0.061	-0.059	-0.052	-0.092	-0.080
	(0.059)	(0.059)	(0.059)	(0.058)	(0.058)
ReliabilityNeu	-0.137*	-0.132*	-0.134*	-0.141*	-0.133*
	(0.057)	(0.057)	(0.056)	(0.056)	(0.055)
TransparencyPos	1.728***	1.735***	1.730***	1.727***	1.734***
	(0.079)	(0.078)	(0.078)	(0.078)	(0.077)
TransparencyNeu	0.777***	0.789***	0.782***	0.781***	0.790***
	(0.068)	(0.068)	(0.067)	(0.065)	(0.064)
Constant	3.292***	2.903***	0.678	3.355***	0.246
	(0.119)	(0.178)	(0.752)	(0.532)	(0.866)
Observations	7800	7800	7800	7800	7800
Number of Subjects	600	600	600	599	599
Estimation Method	Tobit	Tobit	Tobit	Tobit	Tobit
Log PseudoLikelihood	-17169.64	-17127.20	-17068.81	-16924.40	-16749.88
Economic Controls	No	Yes	No	No	Yes
AI Controls	No	No	Yes	No	Yes
Sociodemographics Controls	No	No	No	Yes	Yes

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

it negatively if a technology has a limited number of uses, likely because they themselves expect to make limited use of the technology and therefore only require it to function. The second most chosen dimension is *Data*, selected 57.67% of the time, with positive pairings against all dimensions except *Reliability*. This suggests that individuals value a company that sources its data from public sources rather than keeping everything internal, further indicating a preference for companies that work in collaboration with external entities. We studied which characteristics of a company influence trust in their technology. Our findings indicate that high levels of benevolence and transparency are the most positive factors for subjects to trust a company. This means that having a vision of the benefits of the technology and guiding its development by collecting customer feedback, as well as providing full explanations about the technology, are very positively judged by individuals. It is interesting to note that participants value the company providing explanations about its technology, even if they might not understand these explanations. This suggests that participants potentially believe that society will evaluate the technology, and they will know whether to trust it by listening to the opinions of others. Results also suggest that when communication with society is interrupted, individuals lose confidence in the company, that participants value but do not place as much emphasis on the existence of an institutional framework around the development of the technology, and that participants do not give much value to the quality of data

Table 5. Descriptive statistics of the frequency at which a dimension was preferred to another, for Positive dimension

Variable	Benevolence	Standard	Data	Reliability	Transparency	Overall
Benevolence	X	56.67%	84.33%	76.00%	42.50%	64.88%
Standard	43.33%	X	81.00%	71.17%	36.67%	58.04%
Data	15.67%	19.00%	X	21.50%	11.67%	16.96%
Reliability	24.00%	28.83%	78.50%	X	20.33%	37.92%
Transparency	57.50%	63.33%	88.33%	79.67%	X	72.21%

Table 6. Descriptive statistics of the frequency at which a dimension was preferred to another, for Negative dimension

Variable	Benevolence	Standard	Data	Reliability	Transparency	Overall
Benevolence	X	56.00%	43.83%	41.33%	62.17%	50.83%
Standard	44.00%	X	34.17%	31.17%	50.50%	39.96%
Data	56.17%	65.83%	X	40.67%	68.00%	57.67%
Reliability	58.67%	68.83%	59.33%	X	69.67%	64.13%
Transparency	37.83%	49.50%	32.00%	30.33%	X	37.42%

used to train the technology or the level of performance of the technology. It therefore suggests that participants are mostly concerned by the company providing evidence of its goodwill and assume that AI technologies will be useful to them. This argument is further supported by the reverse logic of the least objectionable characteristics. Individuals are more tolerant of technologies that perform tasks in standard environments and are trained on a reduced dataset compared to other negative characteristics, such as not interacting with society or not operating within a framework. Additionally, it is worth noting that certain sociodemographic characteristics, such as higher altruism and generous strategic thinking, as well as higher levels of knowledge and trust in AI, contribute to increased trust in companies producing AI technologies. This suggests that individuals with more naturally positive dispositions tend to be more trusting of these companies. In conclusion, individuals value demonstrated evidence that a company acts in the interest of social good and are less concerned about technological performance, most likely assuming it will be sufficient for their needs and that they will not need to understand it. These results provide interesting insights for discussions with the general population when introducing AI technology to them.

Acknowledgments

The financial support received from AI Singapore Governance Grant (AISG3-GV-2021-005) is gratefully acknowledged. The usual disclaimers apply. The author(s) declare no competing interest.

References

Aiman-Smith, L., S. E. Scullen, and S. H. Barr (2002). Conducting studies of decision making in organizational contexts: A tutorial for policy-capturing and other regression-based techniques. *Organizational Research Methods 5*(4), 388–414.

Anteby, M. (2010). Markets, morals, and practices of trade:

Jurisdictional disputes in the us commerce in cadavers. *Administrative Science Quarterly 55*(4), 606–638.

Atzmüller, C. and P. M. Steiner (2010). Experimental vignette studies in survey research. *Methodology*.

Bedué, P. and A. Fritzsche (2022). Can we trust ai? an empirical investigation of trust requirements and guide to successful ai adoption. *Journal of Enterprise Information Management 35*(2), 530–549.

Brock, T. C. (1965). Communicator-recipient similarity and decision change. *Journal of Personality and Social Psychology 1*(6), 650.

Cabiddu, F., L. Moi, G. Patriotta, and D. G. Allen (2022). Why do users trust algorithms? a review and conceptualization of initial trust and trust over time. *European Management Journal 40*(5), 685–706.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton university press.

Chiesa, V. and F. Frattini (2011). Commercializing technological innovation: Learning from failures in high-tech markets. *Journal of Product Innovation Management 28*(4), 437–454.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications.* Academic press.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics 14*(4), 583–610.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Frank, D.-A. and T. Otterbring (2023). Autonomy, power and the special case of scarcity: Consumer adoption of highly autonomous artificial intelligence. *British Journal of Management*.

Ghazizadeh, M., Y. Peng, J. D. Lee, and L. N. Boyle (2012). Augmenting the technology acceptance model with trust: Commercial drivers' attitudes towards monitoring and feedback. In *Proceedings of the human factors and ergonomics society annual meeting*, Volume 56, pp. 2286–2290. Sage Publications Sage CA: Los Angeles, CA.

Hengstler, M., E. Enkel, and S. Duelli (2016). Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change 105*, 105–120.

Hooks, D., Z. Davis, V. Agrawal, and Z. Li (2022). Exploring factors influencing technology adoption rate at the macro level: A predictive model. *Technology in Society 68*, 101826.

Larney, A., A. Rotella, and P. Barclay (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes 151*, 61–72.

Lee, J. D. and N. Moray (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies 40*(1), 153–184.

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

Lievens, F. (2017). Organizational image/reputation.

Mayer, R. C., J. H. Davis, and F. D. Schoorman (1995). An integrative model of organizational trust. *Academy of Management Review 20*(3), 709–734.

O'neill, O. (2016). *Justice across boundaries: Whose obligations?* Cambridge University Press.

Ram, S. and J. N. Sheth (1989). Consumer resistance to innovations: the marketing problem and its solutions. *Journal of Consumer Marketing 6*(2), 5–14.

Sternthal, B., R. Dholakia, and C. Leavitt (1978). The persuasive effect of source credibility: Tests of cognitive response. *Journal of Consumer Research 4*(4), 252–260.

Stewart, T. R. (1988). Judgment analysis: procedures. In *Advances in Psychology*, Volume 54, pp. 41–74. Elsevier.

Zuboff, S. (1988). Dilemmas of transformation in the age of the smart machine. *Pub Type 81*, 17.

Appendix

Table A1. Tobit estimations of the impact of levels of dimension on the grade given to vignettes, with selected controls.

	(1) Grade	(2) Grade	(3) Grade	(4) Grade	(5) Grade
BenevolencePos	1.524***	1.531***	1.524***	1.508***	1.512***
	(0.074)	(0.073)	(0.073)	(0.071)	(0.070)
BenevolenceNeu	0.689***	0.695***	0.690***	0.677***	0.677***
	(0.053)	(0.053)	(0.053)	(0.050)	(0.050)
StandardPos	0.619***	0.613***	0.620***	0.624***	0.623***
	(0.050)	(0.050)	(0.050)	(0.050)	(0.050)
StandardNeu	0.473***	0.465***	0.466***	0.476***	0.464***
	(0.046)	(0.046)	(0.046)	(0.046)	(0.045)
DataPos	0.022	0.018	0.018	0.022	0.015
	(0.038)	(0.037)	(0.037)	(0.037)	(0.036)
DataNeu	0.430***	0.429***	0.427***	0.430***	0.427***
	(0.044)	(0.044)	(0.043)	(0.043)	(0.043)
ReliabilityPos	0.140**	0.140**	0.132**	0.149***	0.146***
	(0.044)	(0.043)	(0.043)	(0.041)	(0.040)
ReliabilityNeu	0.123**	0.127***	0.122***	0.124***	0.129***
	(0.037)	(0.037)	(0.037)	(0.035)	(0.035)
TransparencyPos	1.817***	1.809***	1.813***	1.819***	1.814***
	(0.069)	(0.069)	(0.069)	(0.068)	(0.068)
TransparencyNeu	0.873***	0.869***	0.873***	0.873***	0.872***
	(0.046)	(0.046)	(0.046)	(0.044)	(0.044)
SentAmountDG		0.054*			0.048*
		(0.024)			(0.024)
SentAmountUG		0.066*			0.069*
		(0.032)			(0.035)
ScoreAI1			0.143*		0.167**
			(0.064)		(0.061)
ScoreAI5			0.345***		0.379***
			(0.099)		(0.096)
Age				-0.010	-0.020**
				(0.008)	(0.007)
Constant	2.865***	2.433***	0.348	2.631***	-0.301
	(0.101)	(0.160)	(0.732)	(0.520)	(0.820)
Observations	16200	16200	16200	16200	16200
Number of Subjects	600	600	600	600	600
Estimation Method	Tobit	Tobit	Tobit	Tobit	Tobit
Log Pseudolikelihood	-34867.05	-34735.46	-34575.17	-34304.41	-33836.75
Economic Controls	No	Yes	No	No	Yes
AI Controls	No	No	Yes	No	Yes
Sociodemographics Controls	No	No	No	Yes	Yes

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$