Social acceptability of personal data utilization business according to data controllers and purposes

Soichiro Morishita
morishita_soichiro@cyberagent.co.jp
CyberAgent, Inc.
Chiyoda-ku, Tokyo, Japan

Masanori Takano
CyberAgent, Inc.
Chiyoda-ku, Tokyo, Japan

Hideaki Takeda
National Institute of Informatics
Chiyoda-ku, Tokyo, Japan

Faiza Mahdaoui
INP-Toulouse
Toulouse, Occitanie, Haute-Garonne
France

Fumiaki Taka
Kanagawa University
Yokohama, Kanagawa, Japan

Yuki Ogawa
Ritsumeikan University
Kusatsu, Shiga, Japan

ABSTRACT

The rapid development of data analysis technology has enabled various uses of data, including some cases in which data usage cannot be accepted socially. Digital platform operators find it difficult to estimate the reputational risk of releasing new data utilization services. In addition, because the services deployed on the Internet are used across countries, the problem becomes even more complex when the differences between countries are considered. In this study, we conducted a survey in multiple countries, and the analysis takes into account the circumstances among countries to find acceptance of data utilization depending on the entity that utilizes the data. To take into account cultural and legal differences, a survey was conducted in Japan, the United States, the United Kingdom, and France. The number of participants are 782, 509, 235, and 103 respectively. The typology of data utilization is represented by the combination of a “data controller,” “data category,” “processing result,” and “domain (purpose).” We asked the participants about the acceptability of personal data utilization in two stages: whether they believe it is socially acceptable, and whether they would like to use such a service. The former is focused on admissibility, the latter is an intention to use. As a result, we found that acceptability, which is controlled for differences among nations, depends most heavily on an individual’s beliefs about the assurance of the data controller. In short, the individual’s beliefs about the data controller are more important than what the data controller is. Moreover, we investigated the contextual suitability. As a result, we found that the combination of data utilization increased or decreased admissibility and intention to use. In some cases, the intention to use is particularly high even if data utilization is not admitted. This suggests that the service is used out of necessity, including maintaining the status quo.

KEYWORDS

social acceptability, personal data utilization, assurance, orthogonal table, adaptive LASSO

ACM Reference Format:

Soichiro Morishita, Masanori Takano, Hideaki Takeda, Faiza Mahdaoui, Fumiaki Taka, and Yuki Ogawa. 2021. Social acceptability of personal data utilization business according to data controllers and purposes. In *13th ACM Web Science Conference 2021 (WebSci '21), June 21–25, 2021, Virtual Event, United Kingdom*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3447535.3462493>

1 INTRODUCTION

The use of personal data is taking place in a variety of domains. In an essay written by the philosopher Gilles Deleuze in 1992, he presented a vision of a future in which individuals become data[3]. This was an argument that sounded the alarm about so-called *the societies of control*. However, after 9/11, to fight terrorism, the attitudes of citizens toward privacy have been transformed, and they have come to accept societies of control to a certain extent[6, 10]. Now, due to the worldwide COVID-19 epidemic, we are moving in the direction of further acceptance of such societies of control.

However, the use of personal data is not limited to management. During the early 2000s, the use of personal data in e-commerce and advertising became common[15][16]. The use of personal data in the financial (FinTech) and the medical sectors has also flourished[7, 14]. Financial data should be used in the financial field, and medical data should be used in the medical; however, the typology of the data does not necessarily correspond to the classification of the domains in which the data are used. The sharing of medical information for advertising purposes has become a social problem, partly because consent was not obtained [11]. The rapid development of data analysis technology has enabled various uses of data, including some cases in which data usage cannot be socially accepted. It is difficult for digital platform operators to estimate the reputational risk of releasing new data utilization services. Therefore, an investigation of the relationship between the type of data utilization and social acceptability is necessary.

Acceptance varies depending on the entity that utilizes the data. For example, COVID-19 contact tracing applications developed by a private foreign company have been adopted by people in many different countries. By contrast, the French government has adopted

This work is licensed under a Creative Commons Attribution International 4.0 License.

WebSci '21, June 21–25, 2021, Virtual Event, United Kingdom
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8330-1/21/06.
<https://doi.org/10.1145/3447535.3462493>

a centralized system for collecting data, although its usage rate has been low because of concerns about tighter control[8].

We recognized that different contexts, that is, combinations of users, types of data, and domains, will result in different patterns of acceptability, which is our research question in the present study. With regard to the combinations, it is useful to refer to the representation of consent. A consent typology under GDPR was presented in [12]. According to this report, consent is represented with a type of personal data, a type of purposes, a type of processing, and a subject the consent given to. In other words, consent can be expressed as a combination of “data controller,” “data category,” “processing result,” and “domain (purpose).”

Any combination of users, types of data, and domains can result in different patterns and acceptability. A conjoint analysis study of the United States was conducted[9]. The study focused on four factors (status quo, outcome favorability, data sharing, and data protection) and six domains (criminal investigations, crime prevention, citizen scores, healthcare, banking, and employment). As a result, the following three findings are presented. First, outcome favorability and data protection increase social acceptability. Second, there is no preference for the status quo, except in the citizen scores domain. Third, although all technologies have the potential to invade privacy, people have still chosen to accept the technology rather than reject it entirely.

Outcome favorability is assessed against the risk of data leakage and invasion of privacy. People’s perceptions of risk and their relationship with their political beliefs also differ among nations. A study showed the links between perceived threats and political beliefs[2]. This study is an analysis of the World Values Survey, a cross-national survey of social scientists. The results show that economic threats tend to be associated with leftist economic beliefs, violence-related threats tend to be associated with conservative beliefs, and there are few characteristics that can explain the differences among nations.

The rapid development of data analysis technology can give rise to various new and unconventional patterns of data utilization in accord with a variety of national circumstances. In some cases, as seen in the example of the COVID-19 contact confirmation application, data are not used, even if necessary. It is necessary to consider factors related to data protection and the circumstances of each nation.

To obtain knowledge on arbitrary combinations of data use, we use orthogonal tables to enumerate patterns and investigate the acceptability of each.

2 METHOD

A cross-national questionnaire survey was conducted on the acceptance of services according to the type of data utilization, and a regression analysis was conducted using the differences between countries as control variables to find common factors among countries.

2.1 Countries and Procedures

To consider cultural and legal differences, a survey was conducted in Japan, the United States, the United Kingdom, and France. Collaborators were recruited through a crowdsourcing platform. The

platforms used for recruitment and the number of responses obtained are listed in Table 1.

The questionnaire form was created using SurveyMonkey. An attention check[13] was introduced into the questionnaire. In addition, the respondents were informed in advance that they would not be rewarded if they did not pass the attention check. The results of the attention check were automatically judged by SurveyMonkey. If a respondent passed, a non-unique code was issued to the respondent, and the validity of the code was verified by the crowdsourcing platform. In this way, the results of the attention check were judged while ensuring anonymity. Anonymous data collection was conducted in a GDPR-compliant manner with reference to the document by the Committee for Protection of Human Subjects University of California, Berkeley[19].

2.2 Survey Design

The questionnaire consists of three sections. The first section asks about age, gender, and political affiliation. We informed the respondents in advance that there were questions regarding political positions in the explanation and consent form. In the second section, we use an orthogonal table to assign the combinations of “data controller,” “data category,” “processing result,” and “domain (purpose)” that represent the types of data utilization described in the previous section and ask whether the user intends to use the data. In the third section, respondents were asked about their perceptions of “security” for each data controller using a 7-point Likert scale. Each of the above is described in detail below.

2.3 Age, gender, and political affiliation

The ages ranged from 20 to 64 years in increments of 5 years, and 65 years or older. In addition to females and males, we also prepared “others” and no answer as gender options.

Previous studies have shown that attitudes toward public institutions and private companies (foreign capital) differ depending on the political position. In this study, we adopted the left-right ideology scale and asked questions on a 7-point Likert scale. In addition, a “Don’t know” option was provided. If you do not want to answer about your political position, you are instructed to choose “Don’t know”.

2.4 Types of Data Utilization

Based on the aforementioned typology of consent, the typology of data utilization is represented by the combination of “data controller,” “data category,” “processing result,” and “domain (purpose),” as shown in Figure 1. Each item is described in detail below.

2.4.1 Data Controller. The first to be considered are the state (public institutions) and private companies. Private companies are further divided into domestic and foreign. Research institutions have a unique position. The previous studies mentioned above[9] also treated research institutions differently. Based on the above, the user entities are defined as “public institutions,” “research institutions,” “private companies (domestic),” and “private companies (foreign).”

2.4.2 Data Category and Processing Result. The “data category” denotes the classification of the data used for the service, and the

Table 1: Platforms used for recruitment and number of responses received

Country	Number of valid responses	Language of the questionnaire	Platform	Reward
Japan	787	Japanese	Yahoo! Crowdsourcing	115 points (equivalent to 115 JPY)
US	609	English	Amazon MTurk	\$1.86
UK	235			
France	103			

Figure 1: Combination of elements of data utilization

“processing result” denotes a classification of the data derived by the service, both of which are classifications of personal data.

The “data category” and “processing result” are both personal data. According to the World Economic Forum, personal data are classified into the following three categories

- Volunteered
- Observed
- Inferred

In this context, “data category” corresponds to Volunteered and Observed. In this study, we did not distinguish which method was used to obtain the data. The “processing result” corresponds to the Inferred category.

In the design of the “data category” and “processing results,” we decided to provide concrete domain names as representative of the elements of the abstract classification such that the respondents can easily imagine the actual cases in their lives.

For both the “data category” and “processing results,” it is easier for respondents to imagine the situation if the questions are specific. However, if we try to be exhaustive, we may end up with unnecessary questions. However, the number of questions may increase if comprehensiveness is to be ensured. By contrast, if the questions are too abstract, it may be difficult for respondents to imagine the situation. In this study, personal data are broadly classified based on the following two perspectives:

- Internal or External
- Mental or Physical

There are four possible combinations of these:

- Internal and mental: inner feelings (thoughts, beliefs, hobbies, preferences, etc.);
- External and mental: behavior (e.g., speech, behavior, private friendships);
- Internal and physical: physical condition (e.g., health care);
- External and physical: other (social status, income and expenditure assets, public friendships, etc.).

Then, as shown in Table 2, the items representing each quadrant were listed to ensure that the questions were as comprehensive as possible while being specific.

We also considered two additional items. The first item is “personal identification” as a processing result, i.e., which personal data connects the identifier/quasi-identifier. The second item is “appearance” as a data category because it is frequently used in a variety of contexts, although it is also included “other” (external and physical) in the above combinations. For example, appearances are used for personal identification based on facial images and internal/physical status estimations.

2.4.3 Domain. The “domain” should be as specific as the “data category” and “processing result.” Acceptance is a balance of perceived risk and perceived benefit [18], and “domain” is the main factor of benefit and risk of data utilization. In this study, the domains are broadly classified based on the following two perspectives.

- Relative proportions of social benefits and private benefits
- The degree of risk that an individual’s behavior will be constrained

Based on these two perspectives, the following four domains were listed as representatives of each combination.

- Social benefits are relatively large and constraining: Management, control, or monitoring
- Social benefits are relatively large and non-constraining: Decision making of social welfare
- Social benefits are relatively small and constraining: Credit evaluation and insurance screening
- Social benefits are relatively small and non-constraining: Advertisement

As mentioned in the previous chapter, there have long been fears that a control society in which individuals become data will occur in the future. We selected “management, control, or monitoring” as a domain in which social benefits are relatively large and constraining. Next, we selected “decision making of social welfare.” Along with “management, control, or monitoring,” this is important from

Table 2: Correspondence table of data categories and processing result

Combination	Group	Data category	Processing result
Internal and mental	Inner feelings	(Not observable)	Hobby and preference
External and mental	Behavior	Records of Internet activity and location history	Future behavior and intention
Internal and physical	Physical condition	Health care	Illness and life expectancy
External and physical	Other	Earnings, expenses, and assets	Ability to pay and annual income

the perspective of social benefit but is not so constraining for individual behavior. We also selected “credit evaluation and insurance screening” as a domain having private benefits, such as increased life chances, but is considered to have a high degree of restriction on individual behavior. In addition, we selected “advertisement” as a domain in which the benefits are private and the degree of constraint is low, because it is currently widely used for targeted advertising, especially on the Web.

2.5 Answer choices for acceptability

We asked the participants about the acceptability of personal data utilization in two stages. The first stage is whether a participant thinks that this personal data utilization is socially acceptable. The second stage is whether the participant thinks that he/she would like to use a system that adopt this personal data utilization. This two-stage design is used because participants might believe this personal data utilization is irrelevant to their own life [9].

We used the following items:

- I admit that it can exist, and I want to use it.
- I admit that it can exist, but I don't want to use it.
- I don't admit that it can exist, but I want to use it.
- I don't admit that it can exist, and I don't want to use it.

The item “I don't admit that it can exist, but I want to use it,” means that this type of service has already become widespread as a status quo, and there is no choice but to use it. It also corresponds to a situation in which the user is forced to use the service owing to “Abuse of a superior bargaining position by the platform operator,” which became a point of contention in the use of personal data for recruiting in Japan[5].

Because the types of data utilization are expressed by combining the elements shown in the previous section, there may be situations that seem unnatural in certain cases. The following item was added for cases where the respondents felt as such.

- I can't imagine the described data processing activity.

2.6 Trust in the Data Controller

The concept of trust was summarized by Yamagishi *et al.*[20] as shown in Table 3.

Acceptability differs depending on the degree of trust in the data controller. Trust is formed by a combination of those factors presented in Table 3. It is necessary to design a questionnaire by considering such factors separately. In this study, we will focus mainly on “Expectation of the intention of partners” and present other factors to the respondents as prerequisites for the questions to control their responses.

2.6.1 Reliability and Expectation of Competence. Reliability is the expectation for the part that is beyond the intentions and abilities of the partners (including expectations for the order of nature). For example, a belief that a lock will not open by using another key. The expectation of competence is defined as the expectation of a technically competent role performance from those involved in social relationships and systems[1].

The accuracy of the results of data processing depends on the technical limitations and capabilities of the data controller (including data processors commissioned by the data controller). In other words, it depends on expectations of reliability and capability. Users may be disadvantaged if the result of the data processing might be incorrect. In such a case, trust in the data controller is undermined, and the acceptability will be reduced. In terms of the danger of data leakage, acceptability depends on the technical limitations and capability of the data controller. In other words, the cognition of the users regarding technical limitations and capabilities of the data controller are related to acceptability.

Based on the above, the following assumptions were made in the questionnaire:

In all cases, assume that data should be properly managed with sufficient security and the identification of individuals and the estimation of their attributes should be done with sufficient accuracy for each purpose.

2.6.2 Assurance. The expectation for the intention of the partners is the expectation that partners in an interaction will carry out their fiduciary obligations and responsibility. Among such expectations, Yamagishi *et al.* called the “expectation of intentions based on benefits of the partner” an “Assurance.” This is a state in which a violation of one's expectations is detrimental to the partner. For example, contracts are a typical way to create such a state. Even if it does not depend on contracts, reputational risk is also a source of assurance for companies. Reputation is the consistency of a company's image among stakeholders, and when that consistency is lost, the reputation is damaged.

Based on the perspective of reputational risk, in this study, the following three items are listed as factors regarding an assurance.

- The expectation that the partner cares about their reputation and will not do something evil
- The expectation of fairness
- The expectation that partners will give priority to others

The expectation that the partner cares about their reputation indicates the perceptions of participants regarding the data controller's evaluation of reputational risk. The expectation for fairness This

Table 3: Classification of concept of trust

Concept of trust	Reliability			
	Expectation of moral social orders	Expectation of competence		
		Expectation of intention of partners	Assurance	
			Trust	Contextual suitability
				Direct evaluation
				Indirect evaluation (reputation)
				Contextual suitability

is an evaluation of reputation in terms of consistency of corporate image. However, even if the data controller is unfair, the data controller may be trusted if there is an expectation that the data controller will not put others before them. To confirm the difference, we also prepared a question regarding the expectation of not giving priority to others. Based on the above, the following questions were asked for each of the above three items for each user entity using a 7-point Likert scale.

- THEY don't misuse data because they care about their reputation.
- THEY are fair.
- THEY don't always prioritize citizens. (inverted scale)

Note that the string THEY are replaced by "Public agencies," "Research institutes," "Domestic private companies," or "Foreign private companies."

2.6.3 Contextual suitability. As an example of contextual suitability, the following example is given[17].

We unthinkingly trust the mail carrier to bring our mail to the front door. We are immediately suspicious if he or she suddenly starts coming around to the back door or simply walks into the house.

In the framework of this study, the combination of the elements of data utilization refers to the context. The contextual suitability is evaluated by introducing the combinations of two from among "data controller," "data category," "processing result," and "domain (purpose)," as interaction terms of the explanatory variables and confirming their effects. For example, for the interactions between "data controller" and "domain (purpose)," if "public agency : welfare" positively influences acceptability, it can be interpreted that it is contextually suitable for a public agency to utilize data for welfare purposes.

2.7 Variables in Regression Analysis

The variables adopted in the analysis of this paper can be summarized as follows. Note that strings in monospaced fonts are labels for each variable.

- Explained variable
 - `admit`: Admissibility {0, 1} (binary)
 - `use`: Intention to use {0, 1} (binary)
- Explanatory variables
 - Individual attributes
 - * `gender`: Gender {female, male, other, no_answer} (factor)
 - * `age`: Age (normalized for each country)

- * `ideology`: Ideology (7-point Likert scale, normalized for each country)
- * `has_ideology`: Answered regarding ideology {0, 1} (binary)
- Variables that control for differences among countries
 - * `country`: Country {Japan, US, UK, France} (factor)
 - * The interactions "Answered regarding ideology" and Country
- Elements of data utilization
 - * `controller`: Data controller (factor)
 - * `category`: Data category (factor)
 - * `result`: Processing result (factor)
 - * `purpose`: Domain (factor)
 - * The interactions among two of "data controller," "data category," "processing result," and "domain" (factor)
- Items regarding Assurance (depends on Data Controller)
 - `reputation`: Cares about reputation (7-point Likert scale: normalized for each country)
 - `fairness`: Fairness (7-point Likert scale: normalized for each country)
 - `not_others`: Do not give priority to others (7-point Likert scale: normalized for each country)

For a variable that is also a factor, a dummy variable value {0, 1} is generated for each element using one hot encoding. In this section, all factors are used as variables without specifying a reference factor. In this study, all factors are used as variables without specifying any reference factors, because the selection of variables by LASSO sets the coefficient of at least one variable to zero and thus so there is no need to arbitrarily specify a reference.

2.8 Adaptive LASSO

Adaptive LASSO[21], which is employed in this study, is a variable selection method said to satisfy the consistency[4]. Consistency in this context means the consistency of the variables selected (the larger the sample size, the closer the probability of selecting the appropriate variable is to 1) and the consistency of the estimator (the larger the sample size, the closer the regression coefficient is to the true value).

In general, LASSO does not satisfy congruency. However, adaptive LASSO solves this problem by weighting the regularization term as shown in the following equation:

$$\hat{\beta} = \arg \min_{\beta} ||y - X\beta||^2 + \lambda \sum_{j=1}^p w_j |\beta_j|,$$

where y is a vector of values of the dependent variable, X is a matrix of the values of explanatory variables, $\beta = \{\beta_j\}$ is the initial

estimator of the coefficient vector, $\hat{\beta}$ is the resulting coefficient vector, λ is the regularization parameter, and $\mathbf{w} = \{w_j\}$ is the weighting parameter.

The parameter $\mathbf{w}$ is determined as $w_j = 1/|\beta'_j|$ using the consistent estimator β'_j . In general, the result of a multiple regression analysis or ridge regression are used as a consistent estimator. In this study, we use the results of a ridge regression.

3 RESULTS

In this section, we present descriptive statistics and the results of regression analysis of the survey.

3.1 Descriptive Statistics

3.1.1 Individual Attributes. Table 4 shows the gender ratio and the means and standard deviations of age and the left-right ideology scale of the respondents for each country. Regarding ideology, the value of 4 means “neutral.” The respondents of this survey are skewed younger than the actual demographics of each country (Table 4). Because the respondents are registered for crowdsourcing, the sample population of this survey is not representative. However, because this study considers the social acceptability of digital platforms, it is not necessary to consider the bias caused by the adoption of crowdsourcing as a particular problem.

Regarding ideology, the number of “don’t know” responses was particularly high in the US (Table 4). It is possible that the low level of trust in the companies involved in obtaining the data for this survey (CyberAgent, Inc.; Amazon.com, Inc.; and SVMK, Inc.) made people cautious about answering questions about their political beliefs.

3.1.2 Values related to assurance. The distribution of values related to assurance for each country is shown by the box plots in Figure 3–5. The value of 4 means “neutral.”

3.1.3 Acceptance. Cross tabulation of the two explained variables (admit and use) are shown in Figure 6–9.

It should be noted that the data are not representative because this survey was conducted by crowdsourcing collaborators. In particular, the admission rate is higher than that of the general population. In the following analysis, we discuss the relative acceptability among the data utilization factors.

3.2 Results of regression analysis

In this section, we show the results of the regression analysis using Adaptive LASSO described above.

The two dependent variables are applied in different ways. A regression analysis with admit as the explained variable was conducted on all valid responses. We call this experimental pattern “admit.” By contrast, an analysis with use as the explained variable is conducted on the valid responses divided into two parts according to the value of admit. We call the experimental pattern “admit and use” when it is conducted on a set whose value of admit is true, and “reject but use” when it is performed on a set whose value is false.

For the three experimental patterns “admit,” “admit and use,” and “reject but use,” the AUCs were 0.687, 0.610, 0.751, respectively.

3.2.1 Variables that represent differences between countries. The variables encoded using country are integrated as variables that express the differences between nations based on the results of the regression analysis in the previous section.

3.2.2 Decision Tree Analysis. A decision tree analysis (DTA) was conducted on the three experimental patterns described above to determine the dominant factors for each acceptability. The results are shown in Figure 10–12.

3.2.3 Coefficients of explanatory variables and their interaction. The coefficients of explanatory variables and their interaction are shown in Table 5. After variable selection by adaptive LASSO, the regression coefficients were calculated again using ordinary least squares (OLS) with the selected variables. In this study, the regression coefficients with p -value < 0.01 were extracted.

4 DISCUSSION

4.1 Items related to assurance

According to a previous study [20], in a comparison between Japan and the US, Japan is a society built more on assurance than trust. However, according to the results of our survey, as the overall trend for Japan, people did not feel assurance with the data controller (Figure 3–5). At first glance, this appears to be the opposite of the results of the previous study, although the previous study showed that the respondents felt assurance with a specific partner. However, in this study, the question is not about the specific partner but about each type of data controller in general. In a previous study, when asked about trust in general (not in a specific partner), the results showed that Japanese people were generally less trusting than Americans. Therefore, there is no contradiction between the previous study and the results of this survey.

4.2 Discussions based on the result of DTA

In this section, we describe the dominant factors for each pattern based on the results of DTA.

Regarding “admit” (admission that it can exist), assurance to the data controller on reputation is more important than the differences between countries (Figure 10).

Regarding “admit and use” (intention to use with admission), assurance from the data controller about fairness is important (Figure 11).

However, “reject but use” (intention to use without admission) is influenced by ideological tendencies in addition to the assurance from the data controller about the reputation (Figure 12).

In all patterns, the variable indicating the degree of assurance was more dominant than the variable indicating the data controllers. In short, an individual’s belief in relation to the data controller is more important than what the data controller is.

4.3 Discussions based on the results of variable selection

In this section, we discuss the variables shown in Table 5 as a result of variable selection and a regression analysis. As mentioned above, the variables corresponding to each interaction term imply the influence of each element depending on the context.

Table 4: Descriptive Statistics of Individual Attributes

Country	Sample size	Gender				Age		Ideology		
		Female (%)	Male (%)	Others (%)	NA (%)	Mean	SD	Answered (%)	Mean	SD
Japan	782	303 (38.7)	473 (60.5)	0 (0.0)	6 (0.8)	45.6	10.8	677 (86.6)	4.158	0.97
US	609	231 (37.9)	375 (61.6)	0 (0.0)	3 (0.5)	38.2	11.1	406 (66.7)	4.569	2.03
UK	235	83 (35.3)	149 (63.4)	1 (0.4)	2 (0.9)	30.8	9.3	214 (91.1)	3.565	1.44
France	103	30 (29.1)	71 (68.9)	0 (0.0)	2 (1.9)	28.5	8.7	90 (87.4)	3.911	1.51

Figure 2: Distribution of values related to Assurance (Japan) **Figure 3: Distribution of values related to Assurance (US)** **Figure 4: Distribution of values related to Assurance (UK)** **Figure 5: Distribution of values related to Assurance (France)** **Figure 6: Cross tabulation of the two explained variables (Japan)** **Figure 7: Cross tabulation of the two explained variables (US)** **Figure 8: Cross tabulation of the two explained variables (UK)** **Figure 9: Cross tabulation of the two explained variables (France)** **Figure 10: The decision tree of the experimental pattern “admit”** **Figure 11: The decision tree of the experimental pattern “admit and use”** **Figure 12: The decision tree of the experimental pattern “reject but use”**

We evaluate significant levels based on p -value and reference to the variables with $p < 0.1$. In the following explanations, the

lines in Table 5 corresponding to each description are indicated by brackets.

Table 5: This table shows coefficients of explanatory variables and their interaction. See Table 1 and the itemized descriptions in Section 2.7 for what each variable's name means. The items "Individual attributes," "Variables that control for differences among countries," and "Items about assurance" are variables for control. The "Elements of data utilization" are discussed in Section 4.3. Each label corresponds to descriptions shown in the "Labels for description" column.

Variable	Labels for explanation	Admit		Admit and use		Reject but use	
		Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
(Intercept)		-1.351 ***	0.000	-0.616 ***	0.000	-1.177 ***	0.000
Individual attributes							
Age		-0.193 ***	0.000			-0.180 ***	0.000
gender.female		0.714 ***	0.000	0.082	0.112		
gender.male		0.839 ***	0.000			0.368 ***	0.000
ideology		0.185 ***	0.000			0.189 ***	0.000
have ideology				0.326 ***	0.000	-0.022	0.917
Variables that control for differences among countries							
country.US		0.460 ***	0.000	0.182 **	0.006	1.384 ***	0.000
country.Japan		-0.715 ***	0.000	-0.533 ***	0.000	-1.746 ***	0.000
country.UK		0.245 *	0.026			-0.783 ***	0.001
country.France				-0.711 ***	0.000		
have ideology : country.US		0.117 .	0.071			-0.951 ***	0.000
have ideology : country.Japan		0.422 ***	0.000			0.697 *	0.017
have ideology : country.UK		-0.286 **	0.010				
have ideology : country.France				0.699 ***	0.000	-0.664 **	0.007
Items about Assurance							
reputation		0.149 ***	0.000			0.215 ***	0.000
fairness		0.199 ***	0.000	0.173 ***	0.000	0.125 ***	0.000
not_others		0.179 ***	0.000			0.164 ***	0.000
Elements of data utilization							
controller.academic	a-1	0.272 *	0.013				
controller.academic : category.appearance	b-4			-0.362 **	0.009		
controller.foreign_private : category.appearance	a-5, c-5	-0.315 *	0.011			-0.120	0.496
controller.foreign_private : category.behavior	c-4					-0.315 .	0.050
controller.public : category.health	a-2	0.259 *	0.020	-0.001	0.996		
controller.academic : category.health	a-2	0.245 *	0.015				
controller.academic : category.assets	a-6, c-3	0.318 *	0.013	0.140	0.356	0.291 *	0.035
controller.domestic_private : result.identify	c-4	-0.138	0.273			-0.240 .	0.077
controller.foreign_private : purpose.control	c-4	-0.125	0.230	-0.122	0.381	-0.336 .	0.052
controller.public : purpose.welfare	a-2, b-2	0.285 *	0.044	0.200 .	0.060		
controller.academic : purpose.welfare	a-2, c-1	0.157	0.143			0.237 *	0.048
category.health : result.illness	b-1, c-2			0.226 .	0.061	0.505 **	0.006
category.behavior : purpose.control	b-3	-0.137	0.262	-0.267 *	0.044		
category.appearance : purpose.credit	a-5	-0.258 *	0.042			-0.168	0.289
category.health : purpose.credit	a-4	0.211 .	0.059	0.141	0.398	0.223	0.125
category.assets : purpose.welfare	c-1	0.079	0.597			0.304 .	0.065
result.illness : purpose.advertisement	c-4					-0.408 .	0.060
result.interest : purpose.advertisement	b-1			0.305 *	0.038	0.214	0.188
result.illness : purpose.welfare	a-3	0.258 *	0.028				
result.identify : purpose.welfare	b-2			0.243 .	0.065		
result.intention : purpose.welfare	a-3	0.297 *	0.034				

(Signif. codes: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, .: $p < 0.1$)

4.3.1 Admittable or not. When the data controller is a research institution, admit increases (a-1). Originally, the degree of assurance

as people's beliefs are high for research institutions. This result implies that being a research institution increases acceptability.

When public agencies and research institutes handle healthcare data, and their purposes are welfare, it is admitted (a-2). Moreover, when the domain is decision-making regarding social welfare, and the data processing results are illness, life expectancy, future behavior, and intention, it is admitted (a-3). As the reason for this, these combinations are natural in terms of context.

When the processing result is insurance or medical care and the purpose is credit evaluation or insurance screening, it is accepted (a-4).

When appearance is handled, it tends to be rejected, particularly when the data controller is a private foreign company or the domain is a credit evaluation and insurance screening (a-5).

In the case of research institutions handling assets, it was admitted. Moreover, the intention to use is high even if it is rejected (a-6). We discuss this pattern again later in this paper.

4.3.2 Intention to use when admitted. Even if a certain pattern of data use is accepted, it does not necessarily mean that it will be used. However, it tends to be used if it is admitted, rather than if it is rejected. The intention to use when it is admitted represents the conditions under which it is more likely to be used.

When healthcare data are handled and the processing results are illness and life expectancy, the intention to use increases. The same applies when the processing result is hobbies and preferences, and the purpose is advertisement. As in the case of (a-3) above, these combinations are natural in the context of (b-1).

When the domain is decision making for social welfare, intention to use increases when the data controller is a public agency or the data processing result is personal identification (b-2).

When personal identification is performed for management, control, or monitoring, the intention to use decreases (b-3). As mentioned in the introduction, this may be due to people's fear of a controlled society.

When research institutions handle appearance, the intention to use decreases (b-4). Interpreted in conjunction with the results of the previous section, this means that research institutions are generally more acceptable, except when dealing with appearance.

4.3.3 Intention to use when it is not admitted. In the following cases, the intention to use was particularly high even though it was not admitted. This suggests that the service is used out of necessity for reasons such as status quo. For example, in a situation where the purpose is social welfare, the intention to use also increases when the data controller is an research institution and the data category is data on earnings, expenses, and assets (c-1). As shown in Section 4.3.1, both research institutions and welfare were allowed to use personal data. Even people who disapprove of handling such data by research institution unavoidably use the achievement of their researches, provided that the purpose is welfare. With regard to earnings, expenses, and assets, because earnings are examined for public assistance in all countries, people may consider that the handling of their data of income, expenditure, and assets for the purpose of welfare is unavoidable. Moreover, intention to use also increases when the data category is healthcare, and the data processing result is illness and life expectancy (c-2). Because health data contains generally sensitive information, it is natural to want to avoid being managed by third parties, while simultaneously feeling compelled to use the services that manage it. As mentioned above,

it is admitted for research institution handles earnings, expenses, and assets (c-3). The reason for this is uncertain, but for the US, it may be because the Office of Financial Research manages financial data.

In the following situations, it tends to be rejected and not used (c-4).

- A private foreign company handles records of Internet activity and location history
- A private foreign company utilizes data for management, control, or monitoring.
- A domestic private company applies personal identification.
- Illness and life expectancy are handled for advertisements

Some of these situations correspond to specific threats to people. For example, the COVID-19 contact-tracing applications mentioned in the introduction address behavioral history. From the perspective of countries other than the US, The application are developed by a foreign private company. Advertisements related to illness and life expectancy are easily imagined examples and are not comfortable. Conversely, even if there are no concrete examples, people have vague anxiety about management, control, and monitoring by private foreign companies.

It is difficult to explain why people do not especially allow personal identification by private domestic companies. However, this does not indicate that personal identification by private foreign companies is accepted. As mentioned in Section 2.4.2, personal identification is related to appearances, and it tends to be unacceptable for foreign private companies to manage appearance (c-5). Thus, there are complex reasons for the unacceptability of foreign private companies, including appearances.

5 CONCLUSION

In this study, we conducted a survey in multiple countries, and the analysis will take into account the circumstances among countries to find acceptance of data utilization depending on the entity using the data. We asked the participants about the acceptability of personal data utilization in two stages, i.e., whether the participants think that it is socially acceptable, and whether they think that they would like to use such a service. The former is admissibility, the latter is intention to use. According to the results of the decision tree analysis, we found that the acceptability, which is controlled for differences among nations, depends most heavily on an individual's beliefs about the assurance of the data controller. Moreover, as the result of variable selection using adaptive LASSO, the variables corresponding to the data controller are not selected, except for the one corresponding to research institutes. In short, the individual's beliefs about the data controller are more important than what the data controller is. Most of the other factors did not affect acceptability independently, but some interaction terms affected acceptability. We conducted a OLS regression and evaluated the coefficients of interaction terms. We found that the combinations increased or decreased the admissibility and intention to use. Moreover, some patterns showed that the intention to use is particularly high, even though it is not admitted. This suggests that they are using the service out of necessity for reasons such as the status quo.

This research was motivated by the need to understand which data utilization forms would be accepted by society and to reduce

reputation risk. We found that the reputation of the data controller determines acceptability. The meaning of this discovery for digital platform providers is that building trust with users daily would expand business opportunities through the utilization of new data.

In this study, we focused on identifying the factors that are common among nations. A future task is to investigate the differences among nations, which is also important for digital platform providers.

ACKNOWLEDGMENTS

We would like to thank Editage (www.editage.com) for English language editing.

REFERENCES

- [1] Bernard Barber. 1983. *The logic and limits of trust*. Rutgers University Press.
- [2] Mark J. Brandt, Felicity M. Turner-Zwinkels, Beste Karapiriner, Florian Van Leeuwen, Michael Bender, Yvette van Osch, and Byron Adams. 2020. The Association Between Threat and Politics Depends on the Type of Threat, the Political Domain, and the Country. *Personality and Social Psychology Bulletin* (Aug. 2020), 0146167220946187. <https://doi.org/10.1177/0146167220946187> Publisher: SAGE Publications Inc.
- [3] Gilles Deleuze. 1992. Postscript on the Societies of Control. *October* 59 (1992), 3–7. <https://www.jstor.org/stable/778828> Publisher: The MIT Press.
- [4] Jianqing Fan and Runze Li. 2001. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Amer. Statist. Assoc.* 96, 456 (Dec. 2001), 1348–1360. <https://doi.org/10.1198/016214501753382273>
- [5] Kudo Fumiko, Hiromi Arai, and Arisa Ema. 2020. Ethical Issues Regarding the Use of AI Profiling Services for Recruiting: The Japanese Rikunabi Data Scandal.
- [6] J. Holvast. 2008. History of Privacy. In *FIDIS*. https://doi.org/10.1007/978-3-642-03315-5_2
- [7] Shinnosuke Horiuchi. 2019. The merits and demerits of credit risk management based on prediction algorithms: Focusing on the impact on life opportunities and the effectiveness of countermeasures. *Socio-Informatics* 8, 2 (2019), 169–185. http://www.ssi.or.jp/journal/pdf/Vol8No2_12.pdf in Japanese.
- [8] Steve Jacob and Justin Lawaree. 2020. The adoption of contact tracing applications of COVID-19 by European governments. *Policy Design and Practice* 0, 0 (Nov. 2020), 1–15. <https://doi.org/10.1080/25741292.2020.1850404> Publisher: Routledge _eprint: <https://doi.org/10.1080/25741292.2020.1850404>.
- [9] Rabia I. Kodapanakkal, Mark J. Brandt, Christoph Kogler, and Ilja van Beest. 2020. Self-interest and data protection drive the adoption and moral acceptability of big data technologies: A conjoint analysis approach. *Computers in Human Behavior* 108 (July 2020), 106303. <https://doi.org/10.1016/j.chb.2020.106303>
- [10] David Lyon. 2009. Surveillance, power, and everyday life. In *The Oxford Handbook of Information and Communication Technologies*. Oxford University Press, Oxford. <https://doi.org/10.1093/oxfordhb/9780199548798.003.0019> ISBN: 9780199548798.
- [11] Madhumita Murgia and Max Harlow. 2019. How top health websites are sharing sensitive data with advertisers. <https://www.ft.com/content/0fbf4d8e-022b-11ea-be59-e49b2a136b8d>
- [12] Harshvardhan Pandit, Christophe Debruyne, Declan O'Sullivan, and David Lewis. 2019. GConsent - A Consent Ontology Based on the GDPR. 270–282. https://doi.org/10.1007/978-3-030-21348-0_18
- [13] Weiping Pei, Arthur Mayer, Kaylynn Tu, and Chuan Yue. 2020. Attention Please: Your Attention Check Questions in Survey Studies Can Be Automatically Answered. In *Proceedings of The Web Conference 2020*. ACM, Taipei Taiwan, 1182–1193. <https://doi.org/10.1145/3366423.3380195>
- [14] Wullianallur Raghupathi and Viju Raghupathi. 2014. Big data analytics in health-care: promise and potential. *Health Information Science and Systems* 2 (Feb. 2014). <https://doi.org/10.1186/2047-2501-2-3>
- [15] Ben Schafer, Joseph Konstan, and John Riedl. 2000. E-Commerce Recommendation Applications. 5 (Aug. 2000). <https://doi.org/10.1023/A:1009804230409>
- [16] Lee Sherman and John Deighton. 2001. Banner advertising: Measuring effectiveness and optimizing placement. *Journal of Interactive Marketing* 15 (Dec. 2001), 60–64. <https://doi.org/10.1002/dir.1011>
- [17] Robert C. Solomon and Fernando Flores. 2001. *Building Trust*. Oxford University Press.
- [18] Chauncey Starr. 1969. Social Benefit versus Technological Risk. *Science* 165, 3899 (Sept. 1969), 1232–1238. <https://doi.org/10.1126/science.165.3899.1232> Publisher: American Association for the Advancement of Science Section: Articles.
- [19] UC Berkeley Committee for the Protection of Human Subjects. 2020. Mechanical Turk (MTurk) for Online Research. <https://cphs.berkeley.edu/mechanicalturk.pdf>
- [20] Toshio Yamagishi and Hisashi Komiyama. 1995. Significance and the Structure of Trust—Theoretical and Empirical Research on Trust and Commitment Relations—. *INNS Journal* 2 (1995), 1–59. in Japanese.
- [21] Hui Zou. 2006. The Adaptive Lasso and Its Oracle Properties. *J. Amer. Statist. Assoc.* 101, 476 (Dec. 2006), 1418–1429. <https://doi.org/10.1198/016214506000000735> Publisher: Taylor & Francis _eprint: <https://doi.org/10.1198/016214506000000735>.