On the Feasibility of Predicting Users' Privacy Concerns using Contextual Labels and Personal Preferences

Yaqing Yang
Tony W. Li
Haojian Jin

Tsinghua University
UC San Diego

Collective Privacy Norms

Companies should behave under their users' privacy expectations.

Individual Privacy Preferences

Users have **differing levels of sensitivity** to various types of contextual information across domains.

related work (1)

Alan Westin Privacy Segmentation Index

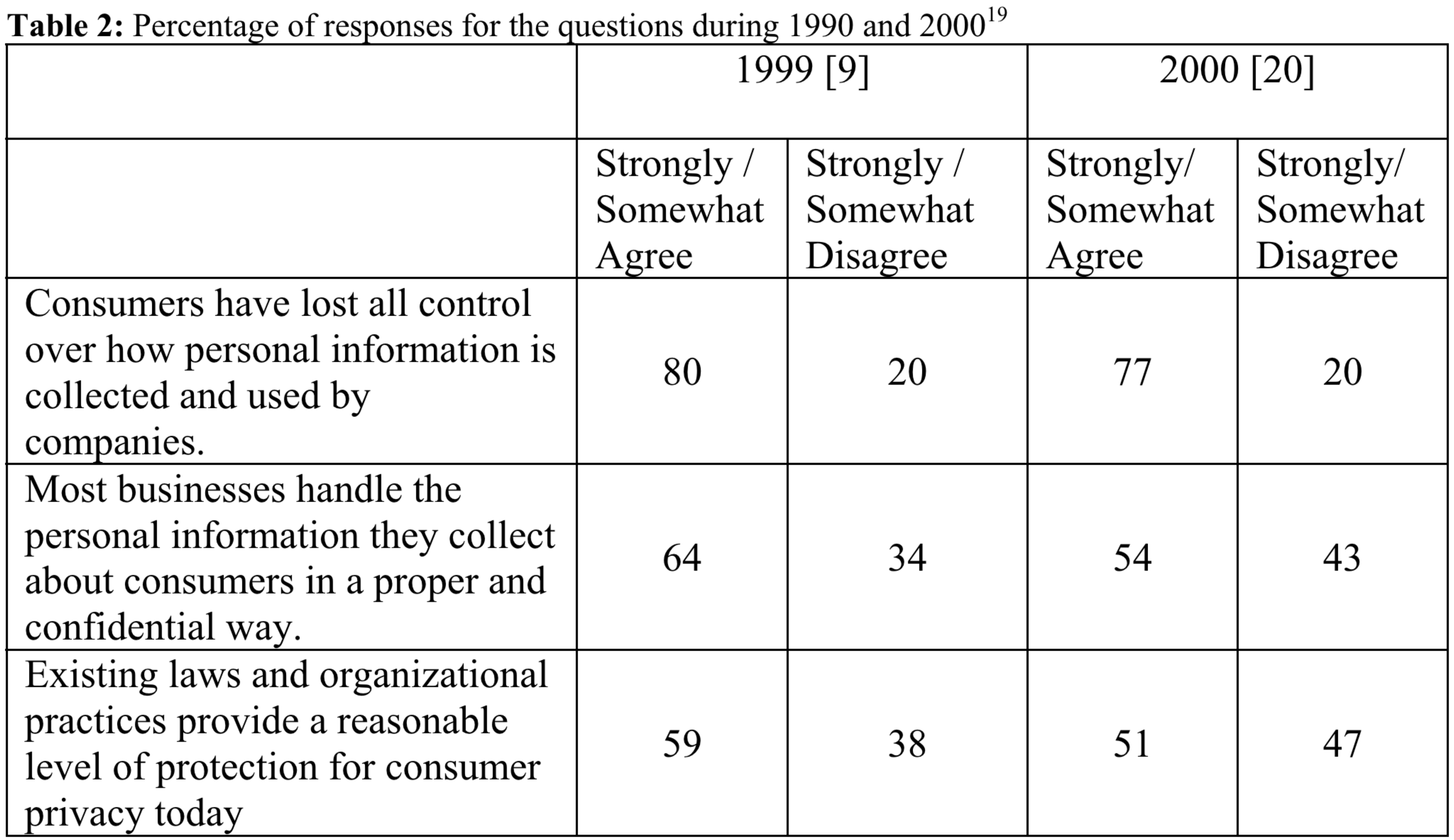


Table 2: Percentage of responses for the questions during 1990 and 2000[19]

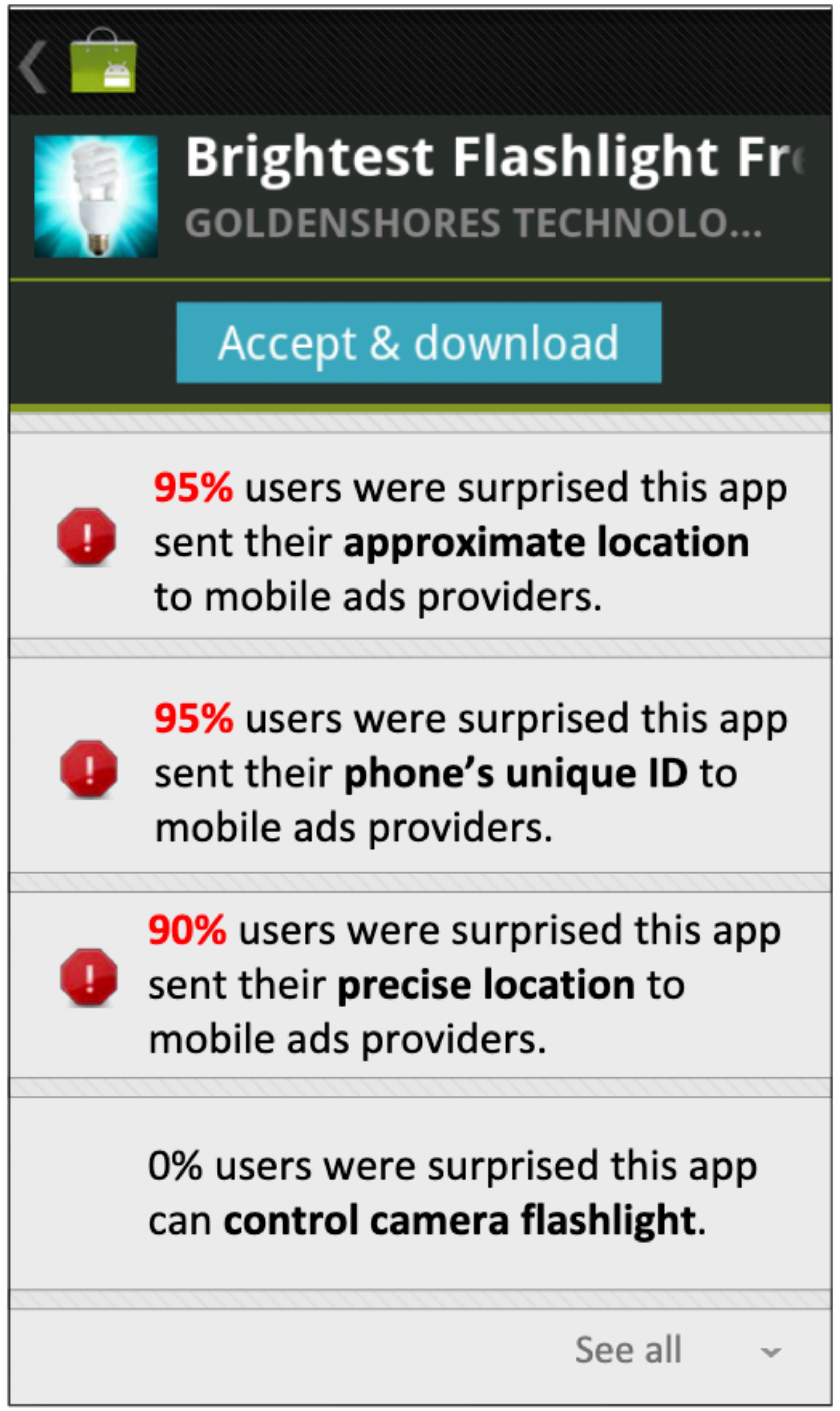
	1999 [9]		2000 [20]	
	Strongly / Somewhat Agree	Strongly / Somewhat Disagree	Strongly/ Somewhat Agree	Strongly/ Somewhat Disagree
Consumers have lost all control over how personal information is collected and used by companies.	80	20	77	20
Most businesses handle the personal information they collect about consumers in a proper and confidential way.	64	34	54	43
Existing laws and organizational practices provide a reasonable level of protection for consumer privacy today	59	38	51	47

- Low correlation
- Not updated since 1995

[1] Privacy indexes: a survey of Westin's studies. 2005
[2] Would a privacy fundamentalist sell their DNA for $1000... if nothing bad happened as a result? The Westin categories, behavioral intentions, and consequences. SOUPS2014

related work (2)

Cluster Similar Users



95% users were surprised this app sent their **approximate location** to mobile ads providers.

95% users were surprised this app sent their **phone's unique ID** to mobile ads providers.

90% users were surprised this app sent their **precise location** to mobile ads providers.

0% users were surprised this app can **control camera flashlight**.

- **Require significant data**
- **Hard to generalize**

[1] Modeling Users' mobile app privacy preferences: Restoring usability in a sea of permission settings. SOUPS 2014
[2] Reconciling Mobile App Privacy and Usability on Smartphones: Could User Privacy Profiles Help?

ContextLabel

modeling **individual** users' privacy expectations by understanding users' **underlying reasoning process** of forming privacy-related opinions.

Two RQs

RQ1: Are users rational towards their contextual privacy concerns?

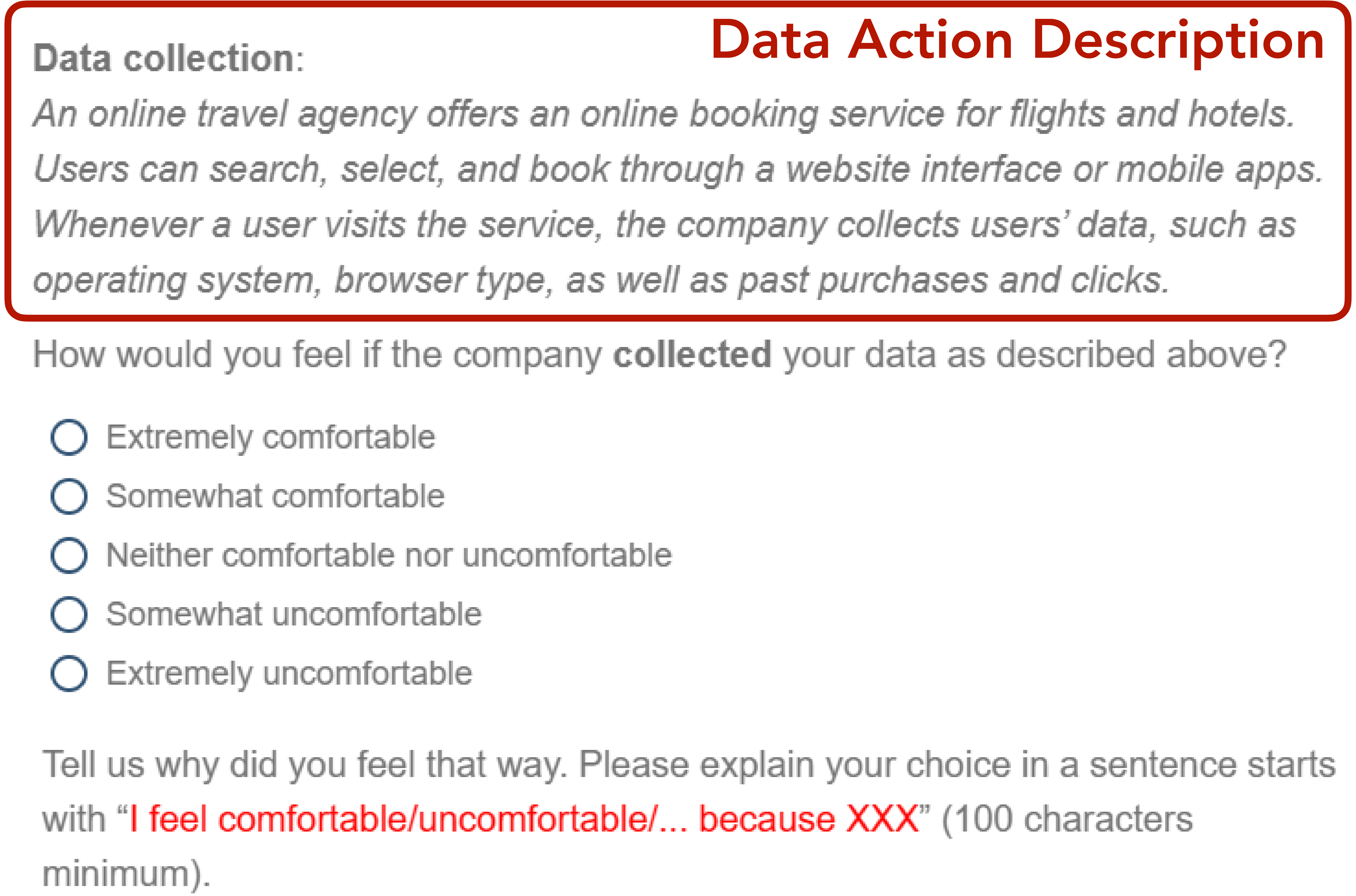
RQ2: How to capture contextual information in privacy scenarios and use it to predict privacy concerns?

RQ1

Are users somewhat rational?

Collecting Privacy Concerns through 5-day Surveys (N=38)

Rate privacy across-domain data actions (N=43) and explain



Data Action Description

Data collection:

An online travel agency offers an online booking service for flights and hotels. Users can search, select, and book through a website interface or mobile apps. Whenever a user visits the service, the company collects users' data, such as operating system, browser type, as well as past purchases and clicks.

How would you feel if the company **collected** your data as described above?

○ Extremely comfortable

○ Somewhat comfortable

○ Neither comfortable nor uncomfortable

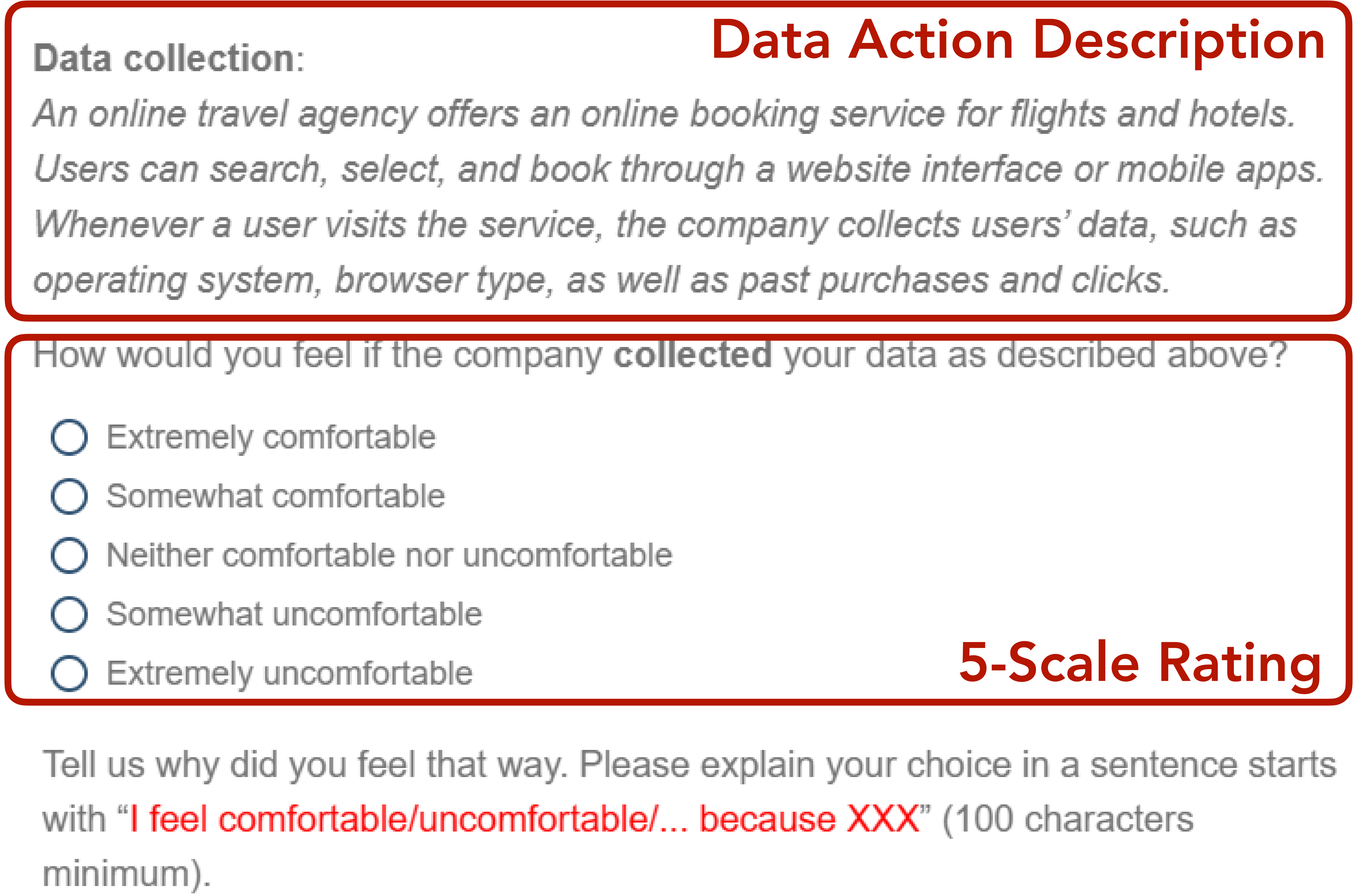
○ Somewhat uncomfortable

○ Extremely uncomfortable

Tell us why did you feel that way. Please explain your choice in a sentence starts with "I feel comfortable/uncomfortable/... because XXX" (100 characters minimum).

Collecting Privacy Concerns through 5-day Surveys (N=38)

Rate privacy across-domain data actions (N=43) and explain



Data Action Description

Data collection:

An online travel agency offers an online booking service for flights and hotels. Users can search, select, and book through a website interface or mobile apps. Whenever a user visits the service, the company collects users' data, such as operating system, browser type, as well as past purchases and clicks.

How would you feel if the company **collected** your data as described above?

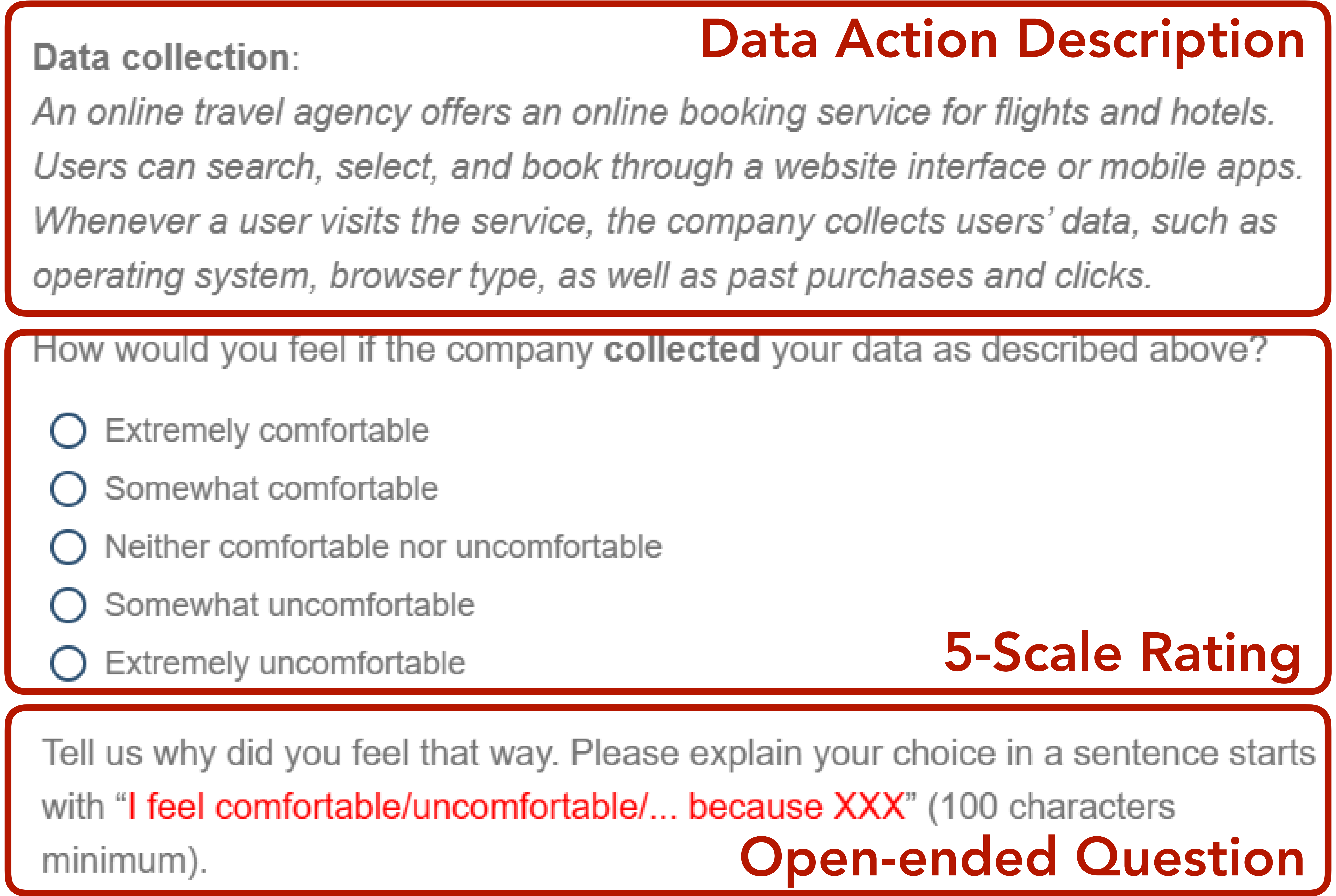
○ Extremely comfortable
○ Somewhat comfortable
○ Neither comfortable nor uncomfortable
○ Somewhat uncomfortable
○ Extremely uncomfortable

5-Scale Rating

Tell us why did you feel that way. Please explain your choice in a sentence starts with "I feel comfortable/uncomfortable/... because XXX" (100 characters minimum).

Collecting Privacy Concerns through 5-day Surveys (N=38)

Rate privacy across-domain data actions (N=43) and explain



Data Action Description

Data collection:

An online travel agency offers an online booking service for flights and hotels. Users can search, select, and book through a website interface or mobile apps. Whenever a user visits the service, the company collects users' data, such as operating system, browser type, as well as past purchases and clicks.

How would you feel if the company **collected** your data as described above?

○ Extremely comfortable

○ Somewhat comfortable

○ Neither comfortable nor uncomfortable

○ Somewhat uncomfortable

○ Extremely uncomfortable

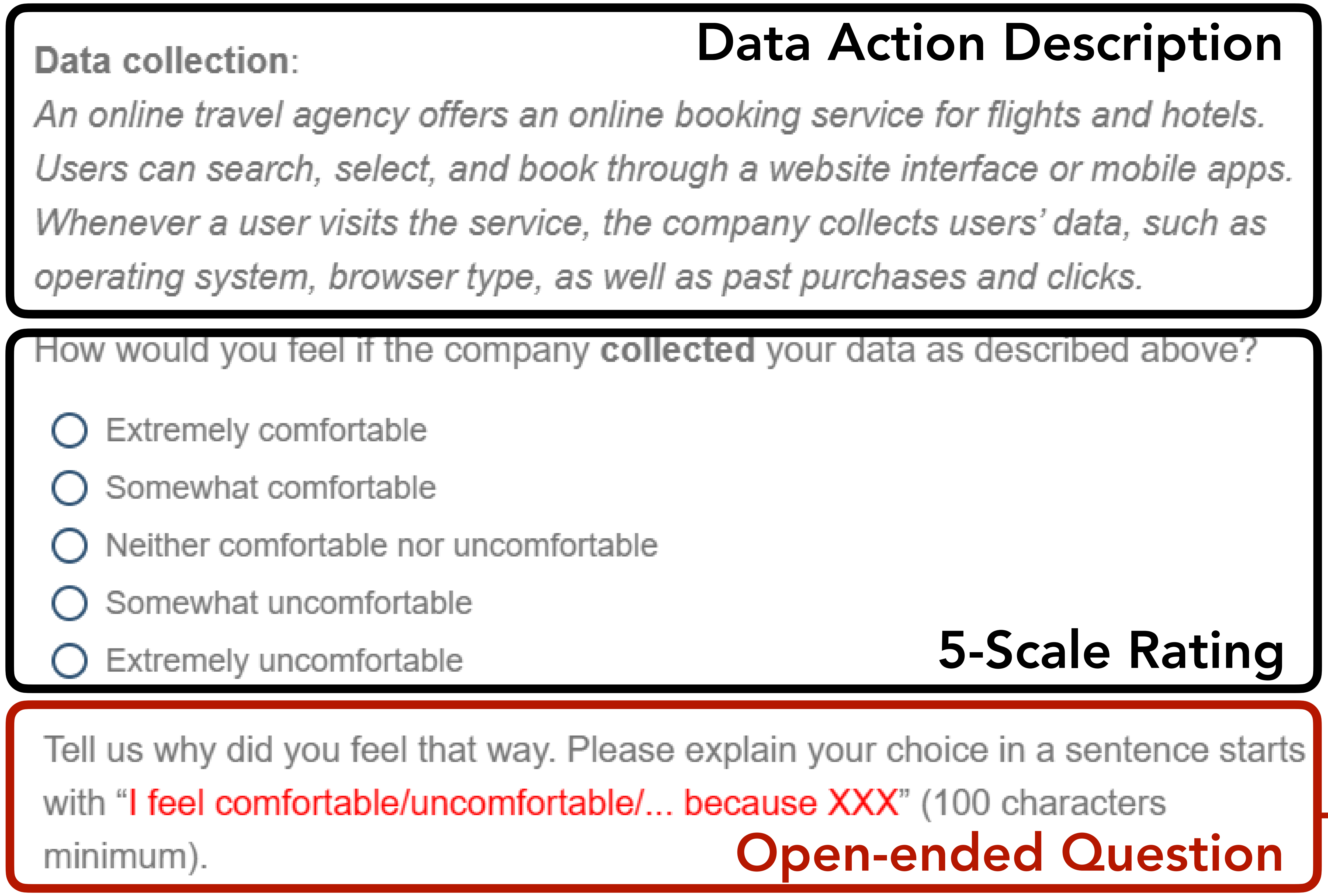
5-Scale Rating

Tell us why did you feel that way. Please explain your choice in a sentence starts with "I feel comfortable/uncomfortable/... because XXX" (100 characters minimum).

Open-ended Question

Collecting Privacy Concerns through 5-day Surveys (N=38)

Rate privacy across-domain data actions (N=43) and explain



Data Action Description

Data collection:

An online travel agency offers an online booking service for flights and hotels. Users can search, select, and book through a website interface or mobile apps. Whenever a user visits the service, the company collects users' data, such as operating system, browser type, as well as past purchases and clicks.

How would you feel if the company **collected** your data as described above?

○ Extremely comfortable
○ Somewhat comfortable
○ Neither comfortable nor uncomfortable
○ Somewhat uncomfortable
○ Extremely uncomfortable

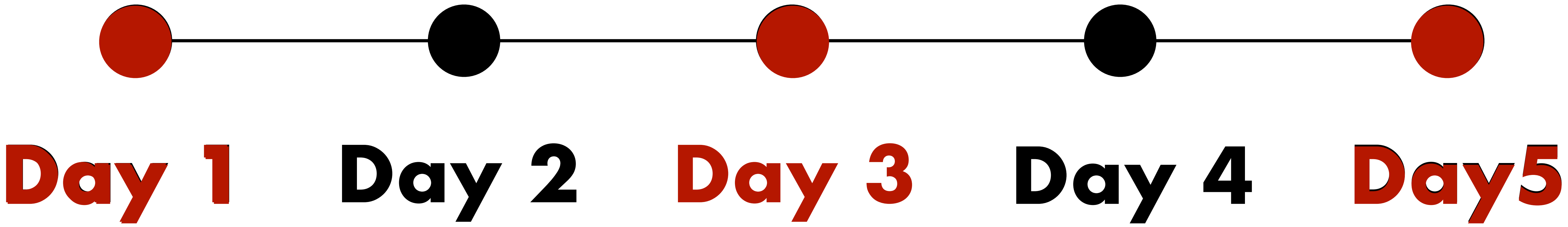
5-Scale Rating

Tell us why did you feel that way. Please explain your choice in a sentence starts with "I feel comfortable/uncomfortable/... because XXX" (100 characters minimum).

Open-ended Question

Crowd workers label free-text explanation using 14 concern categories

Include same 3 data actions and Westin's index in 3 surveys



Day 1 Day 2 Day 3 Day 4 Day5

Check consistency of responses from the same participant across different surveys

How do we know if users are rational?

1. Quantitive
2. Qualitative

Results

Attitudes toward the selected scenario (ICC 0.74) and Westin's index(ICC 0.8) remain consistent.

80% privacy concern categories remain the same.

Results

Inconsistent attitudes are boundary cases.

Action

Based on collected user data, a social App only show users the posts they are likely to engage and hide others.

☺ Day1 ☹ Day3 ☺ Day5

“able to save time…but I don’t like a third party hiding content from people or businesses that I am willingly following”

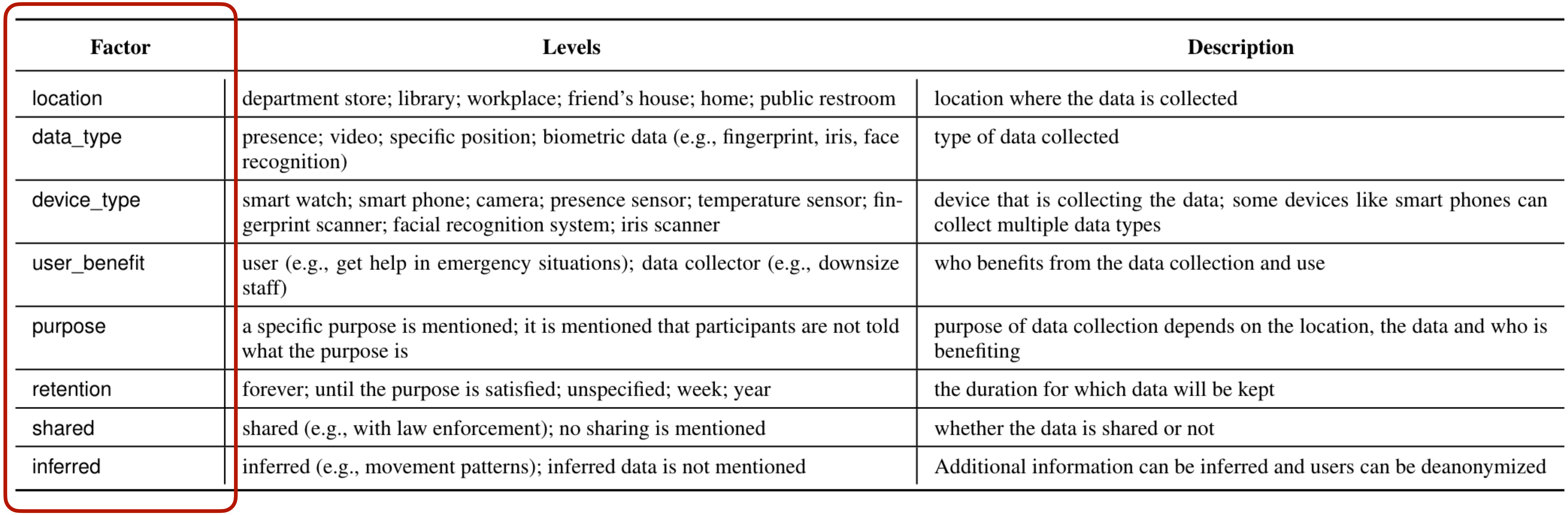
RQ2

How to capture contextual information and use it for privacy concern prediction?

Prior work

Factorial Vignette Surveys & Contextual Factors

categorical factors

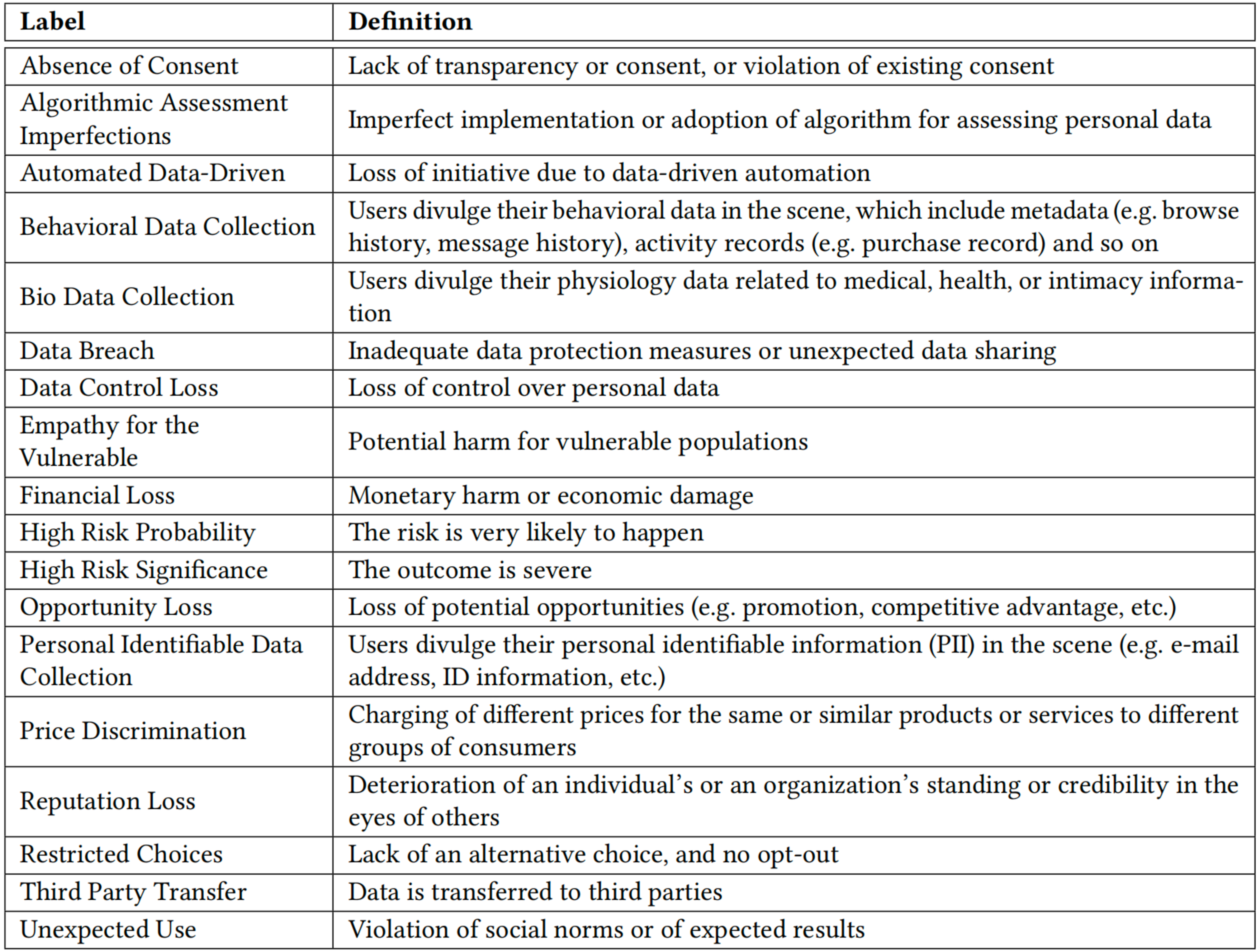


Factor	Levels	Description
location	department store; library; workplace; friend's house; home; public restroom	location where the data is collected
data_type	presence; video; specific position; biometric data (e.g., fingerprint, iris, face recognition)	type of data collected
device_type	smart watch; smart phone; camera; presence sensor; temperature sensor; fingerprint scanner; facial recognition system; iris scanner	device that is collecting the data; some devices like smart phones can collect multiple data types
user_benefit	user (e.g., get help in emergency situations); data collector (e.g., downsize staff)	who benefits from the data collection and use
purpose	a specific purpose is mentioned; it is mentioned that participants are not told what the purpose is	purpose of data collection depends on the location, the data and who is benefiting
retention	forever; until the purpose is satisfied; unspecified; week; year	the duration for which data will be kept
shared	shared (e.g., with law enforcement); no sharing is mentioned	whether the data is shared or not
inferred	inferred (e.g., movement patterns); inferred data is not mentioned	Additional information can be inferred and users can be deanonymized

domain-specific

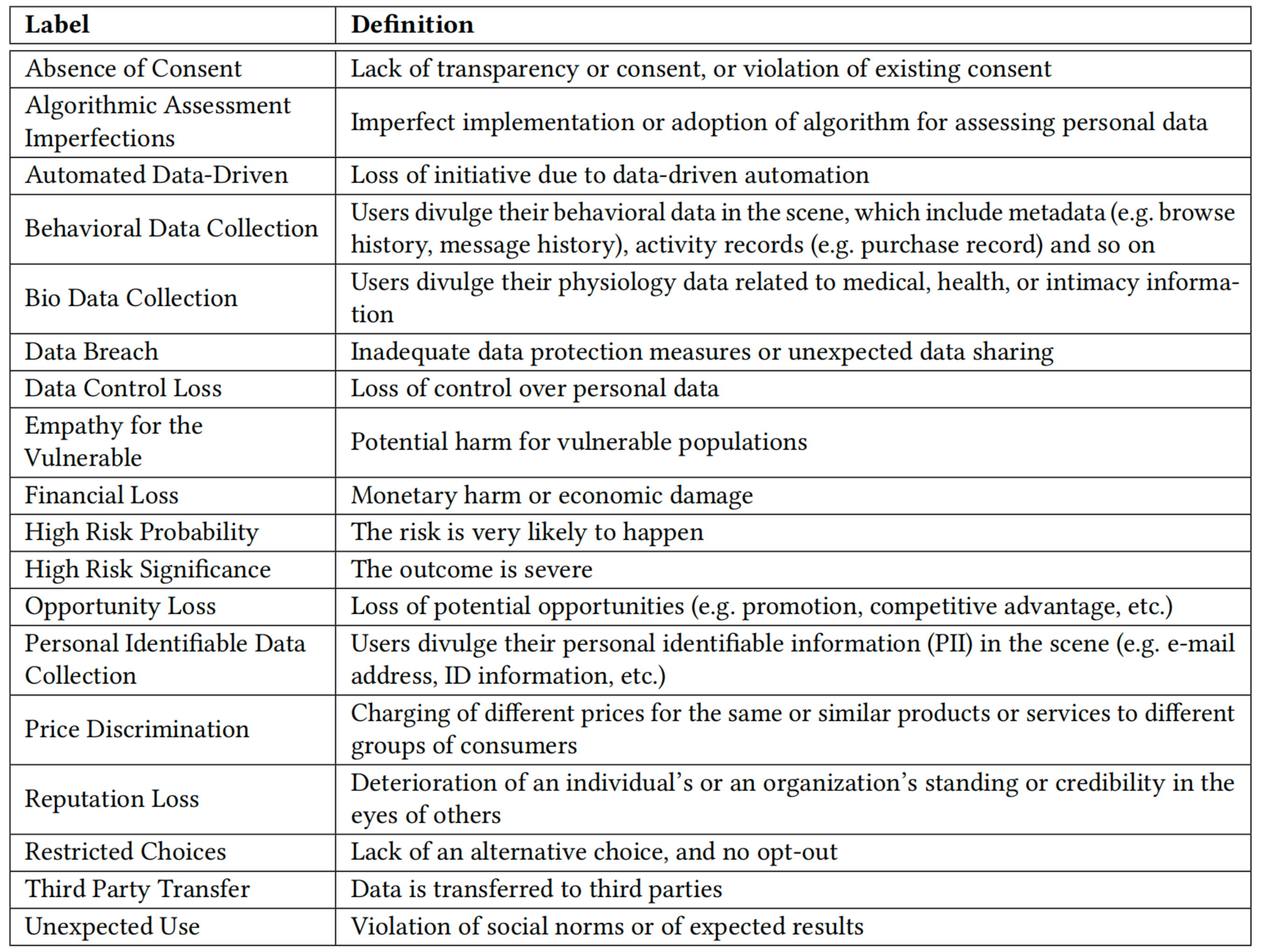
[1] Privacy Expectations and Preferences in an IoT World. SOUPS2017

Non-Exclusive ContextLabels



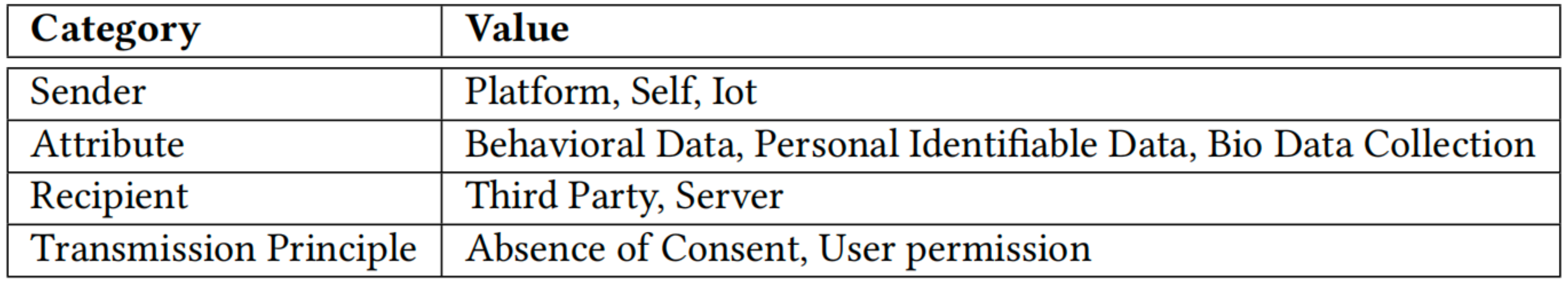
Label	Definition
Absence of Consent	Lack of transparency or consent, or violation of existing consent
Algorithmic Assessment Imperfections	Imperfect implementation or adoption of algorithm for assessing personal data
Automated Data-Driven	Loss of initiative due to data-driven automation
Behavioral Data Collection	Users divulge their behavioral data in the scene, which include metadata (e.g. browse history, message history), activity records (e.g. purchase record) and so on
Bio Data Collection	Users divulge their physiology data related to medical, health, or intimacy information
Data Breach	Inadequate data protection measures or unexpected data sharing
Data Control Loss	Loss of control over personal data
Empathy for the Vulnerable	Potential harm for vulnerable populations
Financial Loss	Monetary harm or economic damage
High Risk Probability	The risk is very likely to happen
High Risk Significance	The outcome is severe
Opportunity Loss	Loss of potential opportunities (e.g. promotion, competitive advantage, etc.)
Personal Identifiable Data Collection	Users divulge their personal identifiable information (PII) in the scene (e.g. e-mail address, ID information, etc.)
Price Discrimination	Charging of different prices for the same or similar products or services to different groups of consumers
Reputation Loss	Deterioration of an individual's or an organization's standing or credibility in the eyes of others
Restricted Choices	Lack of an alternative choice, and no opt-out
Third Party Transfer	Data is transferred to third parties
Unexpected Use	Violation of social norms or of expected results

Capturing Contextual Nuances: Data Action Annotation



Label	Definition
Absence of Consent	Lack of transparency or consent, or violation of existing consent
Algorithmic Assessment Imperfections	Imperfect implementation or adoption of algorithm for assessing personal data
Automated Data-Driven	Loss of initiative due to data-driven automation
Behavioral Data Collection	Users divulge their behavioral data in the scene, which include metadata (e.g. browse history, message history), activity records (e.g. purchase record) and so on
Bio Data Collection	Users divulge their physiology data related to medical, health, or intimacy information
Data Breach	Inadequate data protection measures or unexpected data sharing
Data Control Loss	Loss of control over personal data
Empathy for the Vulnerable	Potential harm for vulnerable populations
Financial Loss	Monetary harm or economic damage
High Risk Probability	The risk is very likely to happen
High Risk Significance	The outcome is severe
Opportunity Loss	Loss of potential opportunities (e.g. promotion, competitive advantage, etc.)
Personal Identifiable Data Collection	Users divulge their personal identifiable information (PII) in the scene (e.g. e-mail address, ID information, etc.)
Price Discrimination	Charging of different prices for the same or similar products or services to different groups of consumers
Reputation Loss	Deterioration of an individual's or an organization's standing or credibility in the eyes of others
Restricted Choices	Lack of an alternative choice, and no opt-out
Third Party Transfer	Data is transferred to third parties
Unexpected Use	Violation of social norms or of expected results

Annotate data actions using ContextLabels



Category	Value
Sender	Platform, Self, Iot
Attribute	Behavioral Data, Personal Identifiable Data, Bio Data Collection
Recipient	Third Party, Server
Transmission Principle	Absence of Consent, User permission

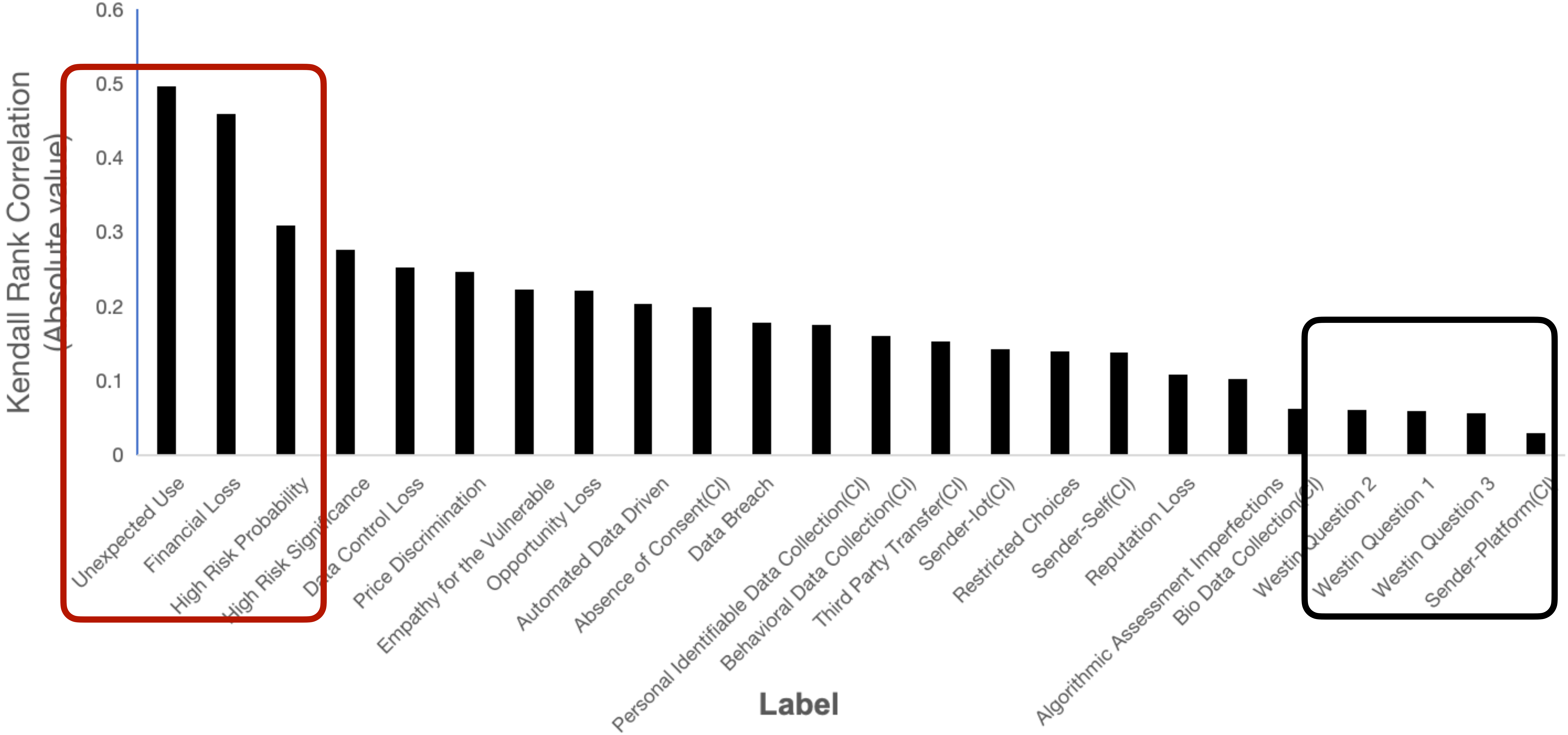
Annotate data actions using Categorical Factors result

Correlation between Contextlabels and Privacy Concerns

Compared to exclusive categories factors and the Westin’s Segmentation Index, non-exclusive contextlabels have stronger correlations with participants’ concerns.

result

Correlation between Contextlabels and Privacy Concerns



Predict Privacy Concerns towards Unseen Scenarios

Multilayer Perceptron Model:

ContextLabel → Concerned / Not Concerned

Model Types:

Global Model: trained on all users' data

Preference Model: trained on personal profile

Baselines:

Westin's Index, Categorical factor

Predict Privacy Concerns towards Unseen Scenarios

ContextLabel has more promising predictive effects

Personal preferences improves prediction

Context Label + Preference achieves best performance

Method	ContextLabel + Preference	ContextLabel	Categorical Factor	Westin's Index
Accuracy	**73%**	64%	59%	56%

Predict Privacy Concerns towards Unseen Scenarios

Multilayer Perceptron Model:

Contextual Label ⟶ Whether user have a specific concern category

Best Performance:

Contextual Label + Preferences, **Acc: 90%**

Takeaway messages

RQ1: Are users rational?

- **Users exhibit a certain level of rationality.**

RQ2: How to capture contextual information and use it to predict privacy concern?

- **ContextLabel can effectively capture contextual information. Combining personal preferences, it can be used for concern prediction.**

On the Feasibility of Predicting Users’ Privacy Concerns using Contextual Labels and Personal Preferences

- Users exhibit a certain level of rationality.
- ContextLabel can effectively capture contextual information. Combining personal preferences, it can be used for concern prediction.

Yaqing Yang Tony W. Li Haojian Jin

Tsinghua University UC San Diego