In the Eye of the Machine

4

CHAPTER 4

Have your ever feared that someone is watching you?

In October 2019, the "Japanese hotel chain, HIS Group . . . apologized for ignoring warnings that its in-room robots were hackable."[1] The hack was revealed on Twitter[2] by "a security researcher [who] warned [the hotel that] the bed-bots [were] easily accessible." This vulnerability allowed "individuals to remotely view video footage from the devices [using a] streaming app." This meant that the in-room robots could potentially have been used to make a candid livestream of a customer's hotel stay.[3]

But camera bots are only a small example of the growing interface between technology and privacy.[4] On one hand, we have computer vision systems, like those embedded in the glasses of Chinese police forces[5] or in public cameras.[6] On the other hand, we have digital records, like those collected by hospitals, insurance providers, search engines, social media platforms, online retailers, mobile phone companies, and voice assistants, such as Alexa or Siri.

What both computer vision systems and data-driven platforms have in common is that they often use the data they collect to train machine-learning algorithms.

This tells us that when it comes to privacy, we need to worry about both the data that can be revealed and the information that can be revealed by models built on this data.

When it comes to data privacy, people are concerned about the possibility of identifying individuals, or gathering sensitive information about groups. When it comes to models, people are concerned about someone learning personal information by interrogating a model. This includes knowing whether a person was part of the data set used to train the model. After all, simply being part of a data set could involve sensitive information (e.g., knowing the mere fact that a person is part of a data set of cancer patients or intelligence agents would reveal sensitive information about that person even if they cannot be pinpointed in the data set).

Reidentification risks are real.[7] A famous story from 1997 involves Latanya Sweeney, now a professor at Harvard but at that time a graduate student at the Massachusetts Institute of Technology (MIT). Sweeney was able to reidentify the medical records of Massachusetts governor William Weld by using publicly available information in an anonymized data set released by the state's Group Insurance Commission.[8] Such reidentification is possible when data entries are characterized by quasi-identifiers, such as ZIP code, sex, and birthdate, that combine to form unique identifiers. Yet quasi-identifiers can also emerge spontaneously from data that has been stripped of any individual characteristics. Consider mobile phone traces. In 2013, a study using mobile phone records found that[9] "in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to . . . the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals."

During the last few decades, scholars have proposed several methods to protect privacy. One of these is the concept of *k*-anonymity, proposed by Sweeney herself.[10] This is the idea that any combination of quasi-identifiers should match at least *k* individuals. But *k*-anonymity has a few problems,[11] since identifying a person within

a group of *k* others can also reveal sensitive information. For instance, in a medical record, we may identify someone within a group of three people who have been diagnosed with HIV, colon cancer, and lupus. Knowing that a person has any of these three conditions constitutes sensitive information. In a real-world example, a fitness company called Strava[12] released an aggregate data visualization of the jogging routes of users of its fitness app, inadvertently releasing information about the location of military bases in Afghanistan. In addition, *k*-anonymity cannot be guaranteed when data sets are combined.[13] For instance, a person who has been to two hospitals that release *k*-anonymous data could be identified by combining these data sets.

These limitations have inspired people to think more creatively about privacy. After all, when we think of attacks on data sets protected by *k*-anonymity, we are thinking about the inferences that a person can make from data. Hence, it is reasonable to think of privacy in terms not only of data, but also of the inferences that we can make from models built on such data. This move from data to models has motivated another approach to data protection, known as *differential privacy*.[14]

In simple terms, differential privacy guarantees that an outside observer cannot know whether a person is part of a data set. This is guaranteed by ensuring that the outcome of any calculations done using the data set does not change—or does not change enough—whether a person is part of a data set or not. As Michael Kearns and Aaron Roth explain in their book *The Ethical Algorithm*:[15] "Suppose some outside observer is trying to guess whether a particular person—say, Rebecca—is in the dataset of interest." If the observer is shown the output of a computation with or without Rebecca's data, "he will not be able to guess which output was shown more accurately than random guessing."

One simple algorithm that can be used to implement differential privacy is *randomized response*.[16] This algorithm, dating back to the 1960s, can be easily explained using an example. Imagine that you want to run a survey to determine how many students in a school use drugs. In principle, drug users may not want to respond to a

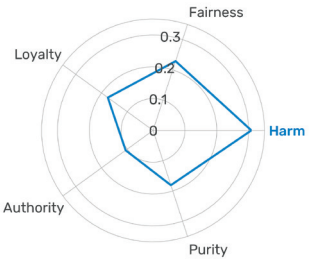
direct question like "Do you use drugs?" because that information could be used against them. Self-censoring would be true even if you promised to keep the data private because the information could be stolen or subpoenaed by law enforcement.[17]

Randomized response offers a solution to this problem by asking people instead to flip a coin (and keep the result confidential), and give true answers only if the coin lands on heads. If it lands on tails, people should flip the coin again and answer "yes" if the coin landed on heads and "no" if it landed on tails. This method helps reveal information, but also gives respondents plausible deniability (since the coin-flip results were never recorded).

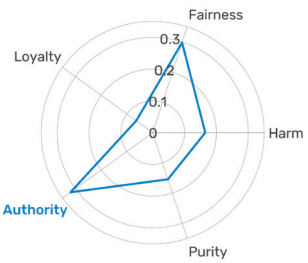
Unfortunately, randomized response is far from bulletproof. It works well if you ask each person to respond only once. But if you ask people to respond multiple times, you become more certain about their true state with each response.[18] Also, even though randomized response works well with helping people reveal sensitive information,[19] it doesn't guarantee trust. In fact, people's trust in the method depends on their ability to understand the procedure.[20] That's why, in recent years, we have seen the rise of more sophisticated privacy-preserving algorithms, such as Rappor, PATE, Federated Learning, and Split Learning.[21]

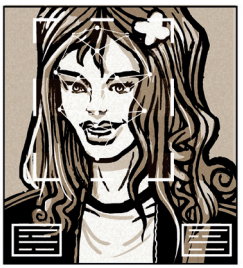
These methods show some of the work that has gone into understanding and protecting privacy in our digital world. But how do these privacy concerns change when the agent behind the data collection efforts is a machine? The reminder of this chapter will be dedicated to exploring people's reactions to scenarios in which people are observed by humans or machines.

Consider the following six scenarios:

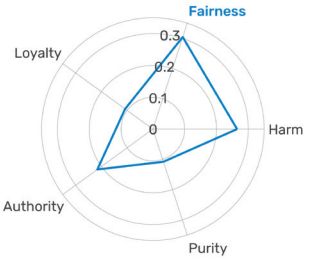


S43 A school is looking to improve the attendance and attention of its students. The school board decides to hire [people/a facial recognition system] to observe students during classes and track the attendance, emotions, and attention of each student.

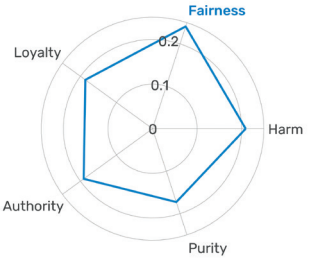




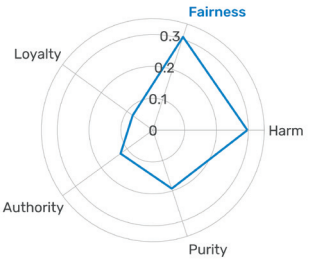
S44 In a city, students are given an ID card that allows them to ride public transportation free of charge. City workers discover that many students are cheating the system by sharing their cards with nonstudent family members. The local government decides to start checking the identity of each rider. To check if the rider's face matches the photo on the ID, an [inspector/facial recognition system] is placed at every access point. The [inspector/facial recognition system] remembers the face of every person checked.



S45 A mall is looking to reduce shoplifting. To improve security, the mall decides to employ a [team of security guards/facial recognition system] to screen everyone who enters or exits the mall. The [team of security guards/facial recognition system] remembers most of the faces screened.



S46 A hotel is looking to build a new poolside bar. The hotel decides to equip the bar with [workers/robots] trained to recognize the face of each guest to keep track of their bills. The [workers/robots] remember everyone they see next to the pool.



S47 An airport management team is seeking to increase security. To do so, the management team decides to equip the airport with [security officers/facial recognition cameras] that will register the face of everyone who enters the airport and track their movements.

We asked people to react to these scenarios by answering the following five questions:

- How **comfortable** are you with this system? (from "Extremely uncomfortable" to "Extremely comfortable")
- How **morally wrong or right** is this system? (from "Extremely wrong" to "Extremely right")
- Do you think this solution **violates** people's privacy? (from" Strongly disagree" to "Strongly agree")
- Would you **recommend** this system to a friend? (from "Would surely not recommend" to "Would surely recommend")
- Would you **use** this system? (from "Would surely not use" to "Would surely use")

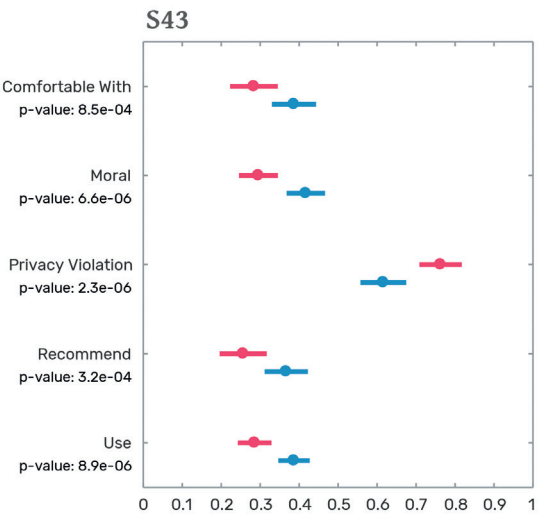


Figure 4.1

Participant reactions to five privacy scenarios: (S43,S44,S45,S46,S47).

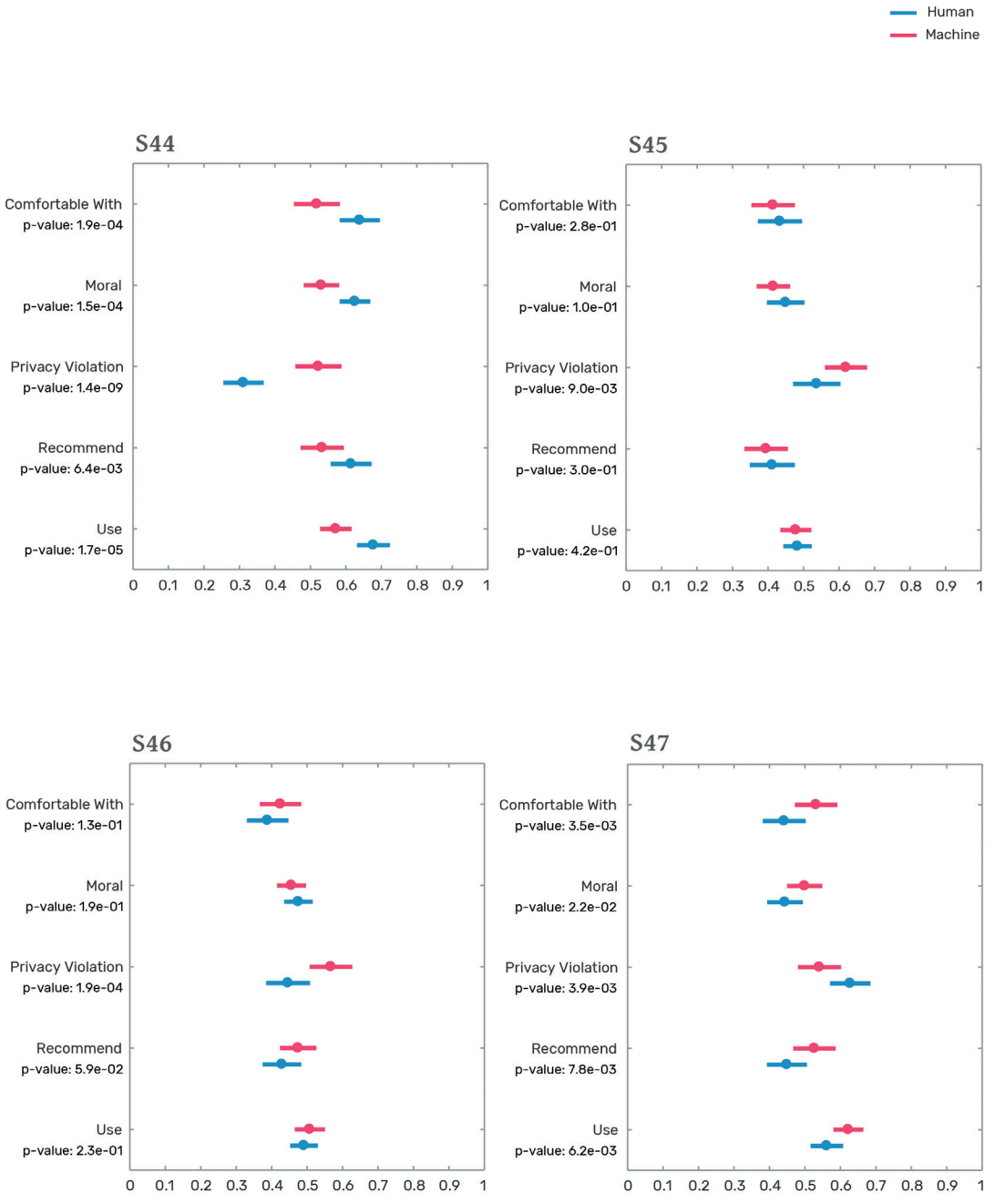
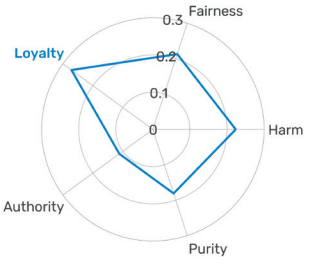

Human
Machine
S44
Comfortable With
p-value: 1.9e-04
Moral
p-value: 1.5e-04
Privacy Violation
p-value: 1.4e-09
Recommend
p-value: 6.4e-03
Use
p-value: 1.7e-05
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
S45
Comfortable With
p-value: 2.8e-01
Moral
p-value: 1.0e-01
Privacy Violation
p-value: 9.0e-03
Recommend
p-value: 3.0e-01
Use
p-value: 4.2e-01
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
S46
Comfortable With
p-value: 1.3e-01
Moral
p-value: 1.9e-01
Privacy Violation
p-value: 1.9e-04
Recommend
p-value: 5.9e-02
Use
p-value: 2.3e-01
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
S47
Comfortable With
p-value: 3.5e-03
Moral
p-value: 2.2e-02
Privacy Violation
p-value: 3.9e-03
Recommend
p-value: 7.8e-03
Use
p-value: 6.2e-03
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

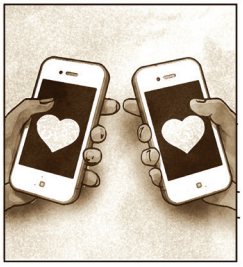
Figure 4.1 presents the result of this exercise. It reveals that people's preference for privacy depends strongly on the circumstances of each scenario. Figures 4.1a and 4.1b show the results for the school monitoring system and the student ID system. In both cases, the respondents have a strong preference for people over machines. They feel more comfortable with human observers and consider human observers to be a more moral choice representing less of a privacy violation. These preferences, however, vanish in the mall security system and the hotel billing system scenarios (figures 4.1c and 4.1d). Here, the preferences for humans and machines are equal. People are mostly indifferent except for the privacy violation dimension, because they consider machine observers to violate privacy more than human observers. Finally, in the airport scenario, there is a clear—albeit mild—preference for machine observers. In this scenario, people report feeling more comfortable and feel that their privacy is less violated when observed by camera systems than by security officers.

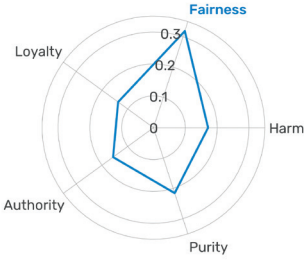
The gradient observed in these five scenarios tell us that people's tolerance of human and machine observers varies by environment. In the student ID and school attendance scenarios, people prefer human observers. This is understandable because people tend to be protective of systems that may violate the privacy of minors. Minors are vulnerable populations who may lack the ability to understand the importance—or lifelong implications—of privacy. At the same time, people are indifferent between human and machine observers in the hotel and mall scenarios, both of which are examples of large commercial settings where people expect some level of private-sector security and surveillance. Finally, we find that the preference for human observers is reversed in the airport scenario, suggesting that people in our sample may be more wary of human observers when they are backed by the coercive force of government.

Next, we explore three scenarios that move away from computer vision systems and involve data-hungry recommender systems. These include predictive purchasing, an online dating system, and a discount travel company.

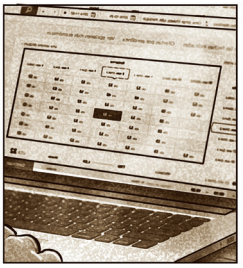


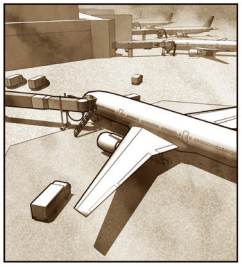
S48 A grocery delivery company announces a new service that uses data on a person's shopping habits to predict the groceries that a person will buy each week. The company assigns to each person a [dedicated shopper/AI digital twin] that uses knowledge of a person's past purchases to predict what groceries to deliver to them.

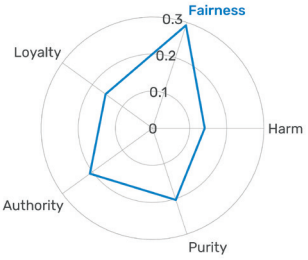




S49 An online dating system announces a new service that uses a person's past choices to set up dates for them automatically. The system requires people to make themselves available one night a week. The system guarantees a weekly blind date for them. The date is set up by a [relationship specialist/AI system].







S50 An online travel company offers a discount vacation system in which users prepay a predefined amount in exchange for letting the system book a discount vacation for them. The company uses [a network of travel agents/AI system] to find and match deals with travelers' preferences.

For each scenario, participants answered the following questions (adapted to each scenario):

- Would you **recommend** the system to a friend?
  (from "Would surely not recommend" to "Would surely recommend")
- Would you **trust** the decisions made by this system?
  (from "Would surely not trust" to "Would surely trust")
- Would you **enroll in/use** this system?
  (from "Would surely not enroll in/use" to "Would surely enroll in/use")
- Have you **ever had** groceries delivered to your home/used online dating sites/used online traveling sites?
  ("No" and "Yes")
- When was the **last time** you had groceries delivered to your home/used online dating sites/used online traveling sites? (from "This week" to "More than a year ago")

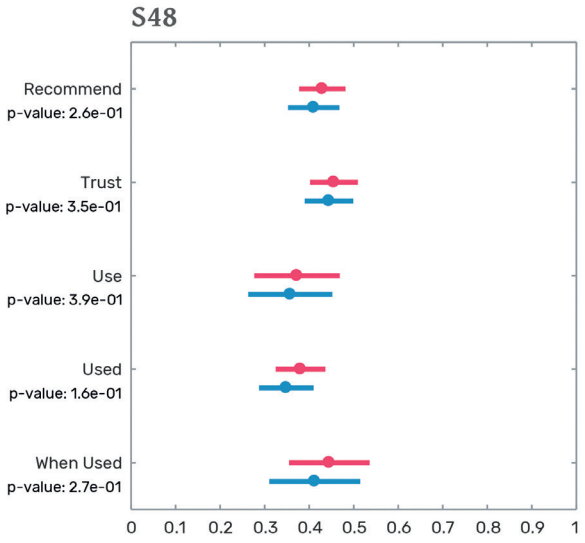
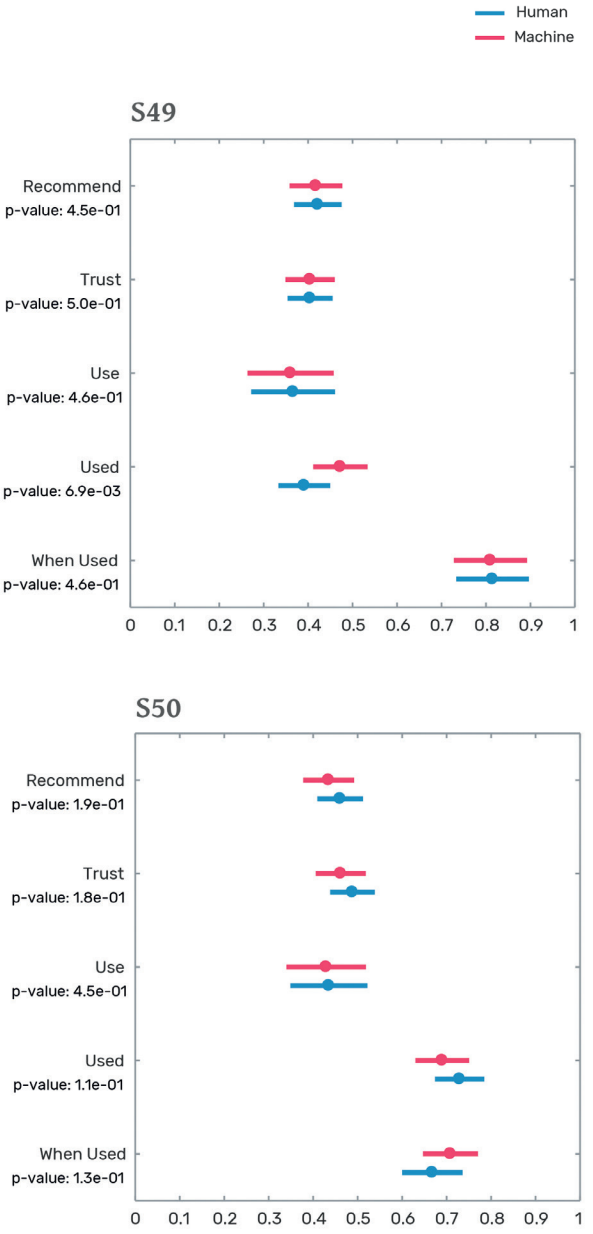


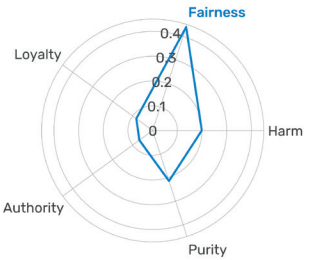
Figure 4.2

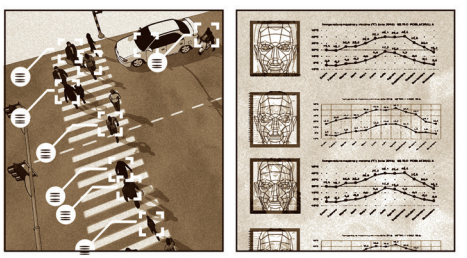
Participant reactions to three recommender system scenarios: (S48,S49,S50).


Human
Machine
S49
Recommend
p-value: 4.5e-01
Trust
p-value: 5.0e-01
Use
p-value: 4.6e-01
Used
p-value: 6.9e-03
When Used
p-value: 4.6e-01
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
S50
Recommend
p-value: 1.9e-01
Trust
p-value: 1.8e-01
Use
p-value: 4.5e-01
Used
p-value: 1.1e-01
When Used
p-value: 1.3e-01
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

In these three examples, we find no strong preference for the use of humans or machines. This is true for systems that enjoy wide adoption today, like online dating, which more than 80 percent of the study's participants report using; or low levels of adoption, such as online groceries, which less than half report using.

Overall, our data suggest some interesting patterns. First, we find a great degree of variation among camera system scenarios. People tend to detest machine observers in scenarios involving schoolchildren and public transportation, but they are indifferent to human and machine observers in private-sector venues. This is echoed by our recommender system scenarios, which are commercial in nature and reveal no big difference between human and machine observers. The only example where we find a preference for machine observers is the airport scenario, which suggests an interaction between human observers and the coercive power associated with governments. To explore that relationship further, consider the following citizen scoring scenario, which was evaluated using the same questions used for the computer vision scenarios.





S51 To improve citizen behavior, a party proposes to implement a scoring system for each citizen. The system is based on [a hotline where citizens can anonymously report others/AI and big data]. The scoring system is used to determine people's creditworthiness, grant admission to public universities, and for hiring and promotion in government jobs.

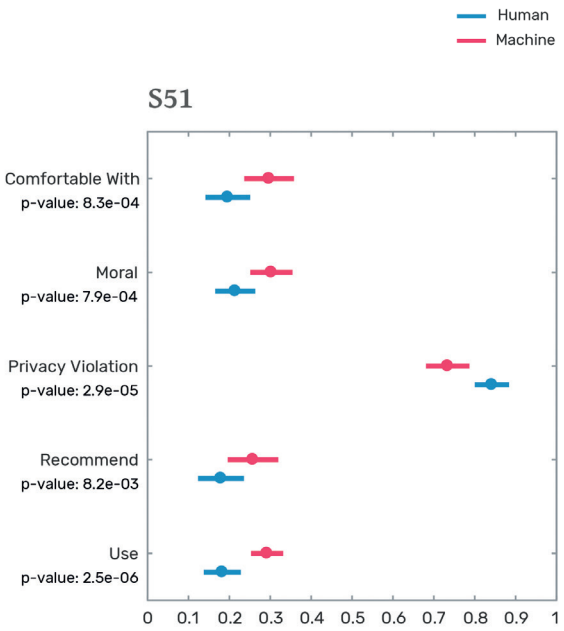


Figure 4.3

Participant reactions to the citizen scoring scenario.

Figure 4.3 shows the results for the citizen scoring scenario. Overall, people reject the idea of citizen scoring, but they do so more strongly when this is implemented in systems that involve people telling on each other than on systems based on algorithms and big data. This aligns with our observation for the airport scenario, but it affords multiple interpretations. On the one hand, having a system where people are incentivized to tell on others has perverse incentives: people could report others not because they've done anything wrong, but because they are rivals or enemies.

Machines are not expected to have such vindictive motives, and hence are less likely to have this perverse incentive. On the other hand, people could be reacting to people telling on others because there are social norms against *ratting out*. Reputation is important to people, and social norms tell us to think twice before we try to ruin another's reputation. Moreover, this scenario—like that of the airport—also involves the coercive power of the state, so this could be yet another interpretation of why people dislike the mechanical approach less.

In this chapter, we compared people's reactions to machine and human observers in a variety of settings. We found that people's preference for human and machine observers varies across scenarios. Yet our results are agnostic about the mental models that people have of machine and human observers. Would people's preference for machine and human observers change if we explicitly described the privacy-preserving protocols involved? Studies about people's attitudes toward randomized response suggest so.[22] But for the time being, this is the end of our journey. In the next chapter, we move away from privacy to focus on another fear induced by machines: the fear of labor displacement.