Group 13 AC221 Project
Trade-offs between public safety and personal privacy in response to COVID-19

Yaoyang Lin Yixuan Di Jingyuan Liu
{*yaoyanglin, yixuan_di, jingyuanliu*}*@g.harvard.edu*

May 2, 2020

1 Introduction

The novel Coronavirus, COVID-19, is spreading with ferocity across a great many countries over the world. Governments, institutions, and individuals are taking drastic measures to flatten the curve of COVID-19 infections, and these measures, while important, may threaten our recently recognized privacy rights. In this project, we will discuss the trade-offs between public safety and personal privacy in the context of the COVID-19 pandemic. We will mainly discuss the data policies' differences in the United States and China, which is a pair of comparative examples. Our goal of this project is to rethink the limits of compromise we could make about privacy usage in this urgent time. The research process includes firstly analyzing the public Coronovirus data in China and the US and evaluating the data privacy condition. We will then discuss the justice of leaking some degree of data privacy for the sake of better control over the epidemic. Lastly, suggestions for data anonymous and policy-making are expected to be provided.

2 COVID-19 Spread and Data Disclosure

The COVID-19, is spreading with ferocity across a great many countries over the world. We have noticed that different countries introduced different policies or regulations to prevent the pandemic, and there seems to exist a trade-off between the epidemic control and citizen's privacy.

2.1 Data disclosure in China

The COVID-19 case is firstly reported in China and the Chinese government implemented strict policies to prevent the virus spread. It not only reports the case number, with the precision of small residential areas, but also mandatorily requires all infected people to report their traces and people touched with in the last 14 days. For example, there's a public data source at GitHub called Wuhan2020[1], which contains infected people's trace records,

[1] https://github.com/wuhan2020/wuhan2020

including family name, gender, age, city, country, street, relation to the virus outbreak place Wuhan, patient id of people get close touched, infected date, hospital, event, transport, etc.

It is acknowledged that the disclosure is of great help to the epidemic control, because healthy people will get alarmed whether they touched infected people or being exposure to dangerous places. The disclosure will also pose moral pressure to people who are not willing to stay at home at this special time. Because if they don't apply the stay-at-home and wear-masks policy, once they get infected, their information will be revealed to the public and get punished by both regulation and public condemnation. Due to the strict policy, the infection in China has been taken under control in two months[2], and the case number is under 85k, much smaller than some other countries like America.

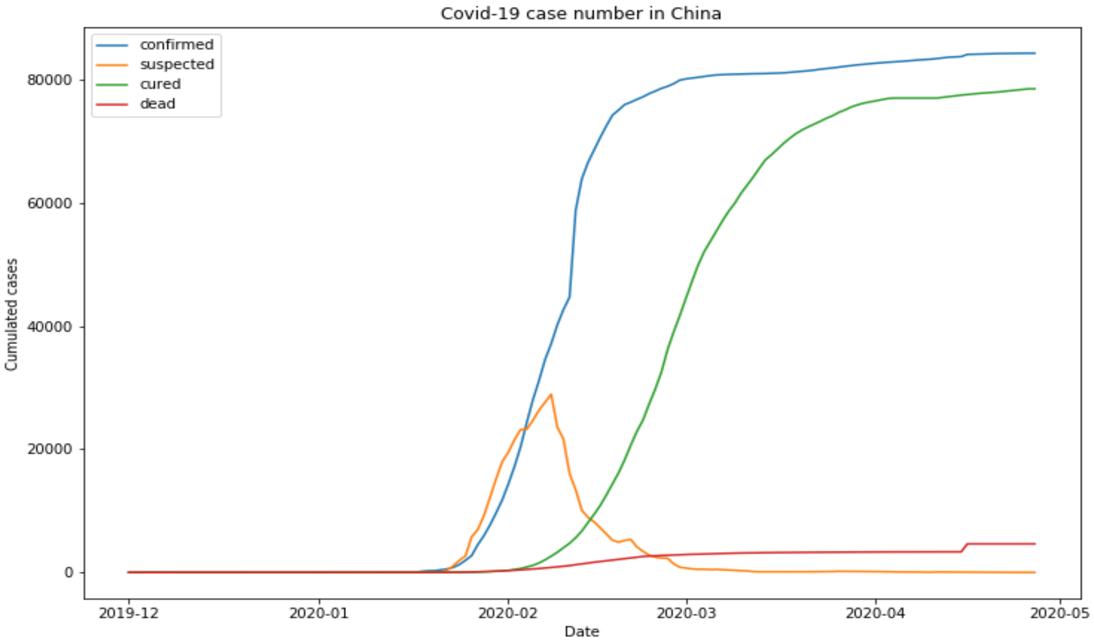


Figure 1: COVID-19 Case Statistics in China

Another interesting example we find is the different degree of data disclosure among provinces in China from the Wuhan2020 patients data. There are four provinces, Anhui, Henan, Hebei and Yunnan have similar number of case records, 200 to 300, but with different proportional of missing column of patients' recent ways of travelling(see table 1). It can be seen that the virus spread speeds are different from figure 2, and a higher rate of missing value correlates to a higher rate of spread speed. Although it is admitted that there are many other factors influencing the spread speed, the relationship can be seen as a heuristic evidence of the importance of data disclosure.

[2] see figure 1 for case number development in China

Province	track case	missing travelling	proportion of missing
Anhui	256	252	0.984375
Henan	233	172	0.738197
Hebei	206	111	0.538835
Yunnan	212	96	0.452830

Table 1: Track case number and missing proportion

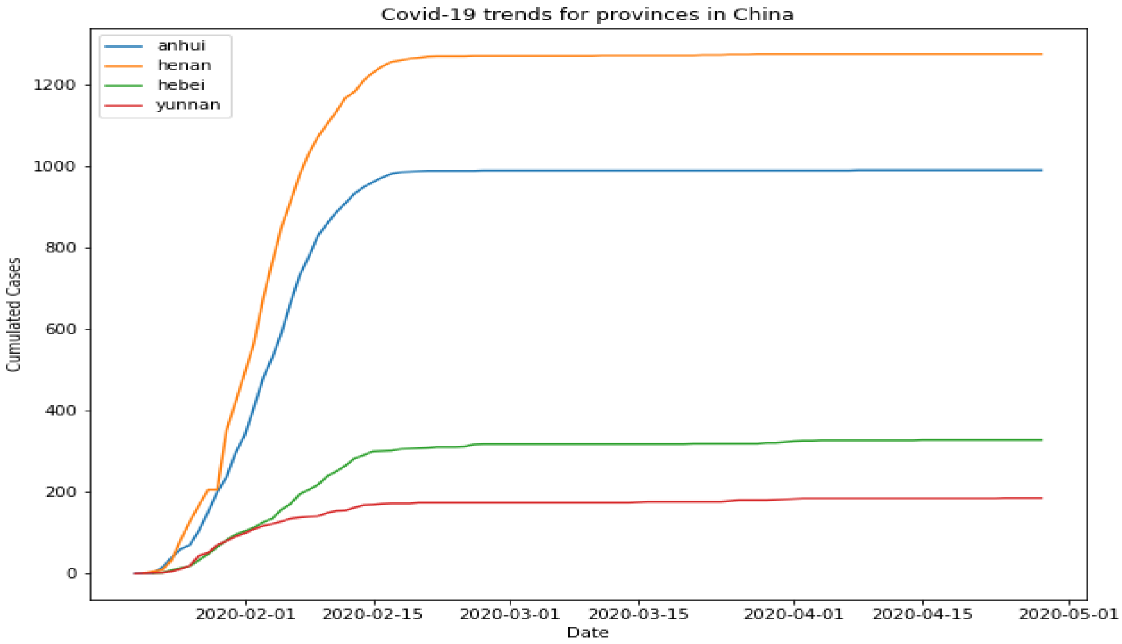


Figure 2: Epidemic spread comparison in four provinces

However, concerns about data privacy have been raised along with the epidemic prevention period. People are questioning that the data disclosure is an infringement of citizen's right of privacy. Take the Wuhan2020 data as an example, with the quasi-identifiers information, age, gender, address, transport, etc., one can easily identify a person, and can get some sensitive information like their living habits, their family address, their ages, and their social circles. In fact, we doubt that much sensitive information is not necessary for epidemic control. While we acknowledge the importance of information disclosure during the special time, the question is whether this degree of disclosure is necessary, and how to find a balance between patients data privacy and epidemic control.

2.2 Data disclosure in the United States

Since the first confirmed case on the 22th Jan 2020, the Coronavirus has spread across the America in an extremely crazy speed. Up till 26th April, there has been 987k confirmed case, and 55415 deaths resulted from COVID-19 (see figure 3 for case number trend in the US)[3].

The CDC, including states governments, has implemented a series of data policies, to reveal the infection number and patients conditions. However, the data disclosure level is significantly lower than that in China. For most of the states, the data privacy is protected well enough to ensure patients not be identified. A typical way of patients data disclosure is the government only announce the number of infected people with the precision of country and city wide. Rather than making public each patients age, gender and other information, they publish the distribution data of all patient groups, or use some generalization methods like group the ages. However, we still find one state, Georgia, disclosed some questionable data. The state government released the death case data, which contains death case's personal information, like exact age, gender, and their living countries. We find that from these quasi-identifiers, there are 630 records out of 895 in total could be uniquely identified.

While the United States have a strong sense of protecting data privacy, we doubt that whether it is wise to protect personal privacy to such an extent that it will actually hinder the pandemic control, infringe other people's right, and cause worse consequence to the whole society. First, citizens have the right of health and safety, the protection of patients' condition and trace may harm other people's right to know where is the dangerous place and know the truth. Everyone needs to know whether they have been exposed to a dangerous situation with a high probability of getting infected. Second, as we all see, the COVID-19 virus has been spread at a crazy speed in the US. If the governments release more data and increase the transparency, people will get alarmed and it may well help preventing the virus spread. Third, the failure of controlling the pandemic will damage the whole country in numerous aspects, such as the economy, development, community safety, people's mental health, and citizen's trust to the government. And these effects will be more serious comparing with the potential harm brought from the invasion of privacy. What we hope to see, is that there should be a way to protect people's privacy while increasing the transparency at the same time.

2.3 Data disclosure in other areas

Different countries and areas have distinct level of data disclosure [6]. Generally, a country or area with stronger centralized power will have tougher control policy to prevent the epidemic, and more disclosure of case data. Table 2 shows some typical patients data disclosure policy in Singapore, Hong Kong(China), South Korea, United Kingdom, Germany, and New York(the US) [5].

[3] Data Sourse: https://raw.githubusercontent.com/canghailan/Wuhan-2019-nCoV/master/Wuhan-2019-nCoV.csv

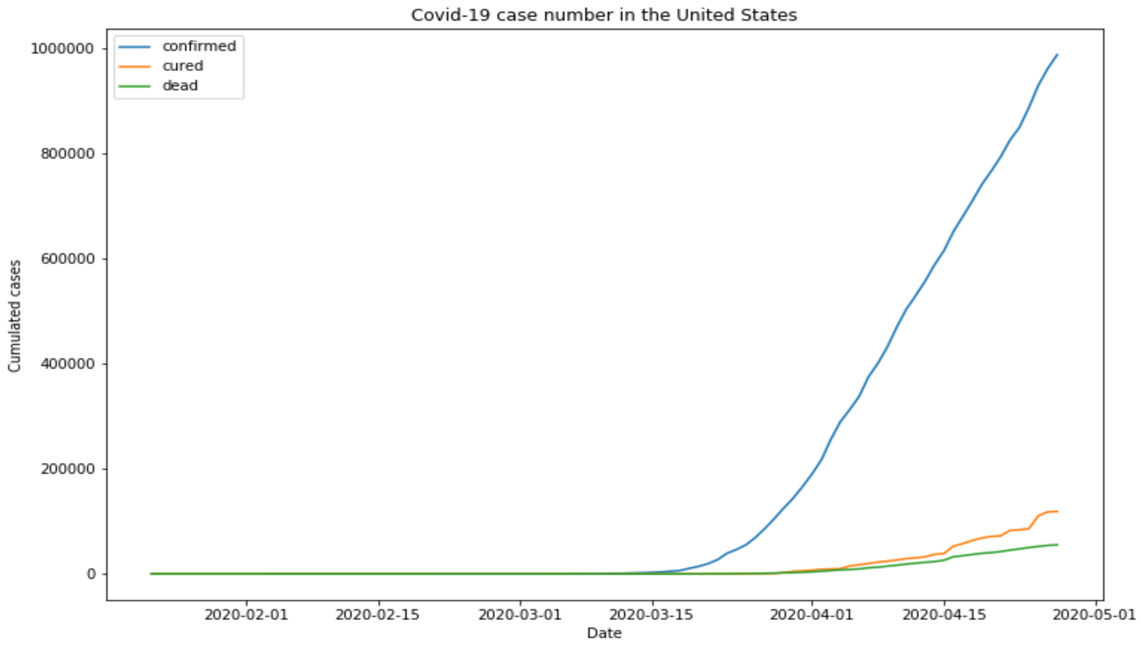


Figure 3: COVID-19 Case Statistics in the US

Disclosure info	Singapore	HK	Korea	UK	Germany	NY
age and gender	✓	✓	✓	✗	✓	✓
travel history	✓	✓	✓	✓	✗	✓
business address	✓	✗	✓	✗	✗	✗
home address	✓	✓	✓	✗	✗	✗
relation to previous cases	✓	✓	✗	✗	✓	✗
nationality	✓	✗	✓	✗	✗	✗
treatment place	✓	✓	✓	✓	✗	✗
travelling places before diagnosing	✓	✗	✓	✗	✗	✗
close contact people	✗	✓	✓	✗	✗	✗
how to get diagnosed	✗	✗	✓	✗	✗	✗
area	✗	✓	✗	✗	✓	✓

Table 2: Data Disclosure policy in different areas

3 Tracking and Public-private Partnerships

In the context of the COVID-19 outbreak, many countries would rely on location tracking to map the evolution of the virus and plan responses. Some governments even deploy digital surveillance tools to impose social control. The government and tech companies have been working together to develop technological solutions: app, websites etc. to fight the COVID-19 outbreak. However, these come with a number of concerns: First, those policies, often developed without appropriate safeguards, will greatly impair people's ability to keep their health conditions private. Second, previous uses of phone records and location data in humanitarian responses were shown to be inefficient and ineffective[4]. Third, placing a country's entire population under monitoring and surveillance would increase the risk of catastrophic eavesdropping in the future.

3.1 Monitoring policy in China

The Chinese government has cooperated with one of the largest tench company Alibaba developing an app, Alipay Health Code, which could analyze their personal data and sort individuals into color-coded categories corresponding to their health status and level of risk for COVID-19. [1]

At checkpoints throughout the WuHan city, police and security guards demanded people present a QR code on their mobile phones to make sure whether they should be allowed to enter subways, malls and other public places. A green code enables its holder to move about unrestricted. Someone with a yellow code may be asked to stay home for seven days. Red means a two-week quarantine. The system is already in use in 200 cities and is being rolled out across the country. Intensive use of the health code is part of the efforts by authorities to revive China's economy while preventing a spike in infections as workers stream back into factories, offices and shops.

Meanwhile, concerns[8] are rising about whether this temporary measure will become a permanent fixture, setting a template for new forms of automated social control that could persist even after the epidemic subsides. Sharing personal data with the authorities further erodes the line between Chinese tech giants and the government. In addition, neither the company nor Chinese officials have explained in detail how the system classifies users. This causes fear and confusion among those who are forced to isolate themselves without knowing why they have red codes. However, based on the interview of Chinese people, most of them think the reward worth the risk: "If we had to use it indefinitely, that would be crazy — just way too big a pain. But for the epidemic, it makes sense."

3.2 Monitoring policy in the United States

Compared to China, an authoritarian regime with a long history, the U.S. who focuses more on democracy and freedom is more careful and hesitant to enable contact tracing or enforced isolation.

Under the great pressure of the economy caused by the epidemic, Google and Apple are working together to add coronavirus tracing to Android and iOS[7]. The basic system uses the phone's Bluetooth to anonymously track who you have been close to. It will be mid-May before the first true rollouts begin but it seems that it tries to facing the challenge of privacy and trust.

[4] CIS-India.Ebola:Abigdatadisaster, 2016.https://cis-india.org/papers/ebola-a-big-data-disaster

To tackling privacy problems, Google and Apple said the app will require explicit user consent to start tracking, and the user can always turn it off—either permanently or temporarily. And the companies are in part responsible for the outcomes, including potential abuses as well as medical successes. Technically, the new system does not collect any true location data: it's all proximity data gathered by Bluetooth using randomly generated, which makes it harder to tie you to any of the other data your phone may carry about you. In addition, the service will only be available to government public health agencies directly involved in coronavirus tracking, a group that Apple and Google say will be small and closely monitored. What they have so far is a blueprint for something promising, but the actual implementation will make the difference. And the world will be watching to see what that means.

During the epidemic, American laws also recognize that extraordinary circumstances require extraordinary measures. For example, New York passed a law this month giving Andrew m. Cuomo unrestricted powers, which could overturn any local regulations, to enforce executive orders during pandemics. However, people are worried about that "It's easy to get into a situation where we empower local, state, or federal governments to take action against a pandemic that fundamentally changes the scope of civil rights in the United States."

3.3 Monitoring policy in other areas

For other typical countries:

- In Singapore, there is a smartphone app, TraceTogether, to help authorities find people who may have been exposed to the virus[9]. It allows people to voluntarily share their information and tracks other people with whom they come in contact via BlueTooth. If any of the app's users contract COVID-19, all users who have come in contact with the said person are notified, along with the government. A government official said the app protected privacy by not revealing their identities to each other.

- In South Korea, government agencies are using surveillance footage, location data from smartphones, and credit card purchase records to help track the recent movements of patients with coronavirus and establish chains of transmission. In the highly networked society, cybermobs use patient data disclosed on government websites to identify patients by name and track them down. The officials are improving data sharing guidelines to minimize patient risk because the invasion of privacy might deter citizens from being tested for the virus.

- In Lombardy, Italy, authorities are analyzing location data from citizens' mobile phones to determine how many people are complying with the government's blockade.

- The U.K. also has plans for its own BlueTooth contact tracing app, and again this would be governmental. The app is used to "enforce social distancing", warning people if they get too close to others or if they spend too much time outside or away from their homes.

- In Israel, the internal security service is preparing to start using caches of mobile phone location data – initially used in counter-terrorism operations – to pinpoint people who may have been infected with the virus.

4 Recommendation

In the background of the unprecedented health crisis, controlling the spread of diseases and protecting public health will undoubtedly become a higher priority. Some sacrifice of privacy is reasonable, but it must be premised on having clear and transparent rules. The government should propose a plan that explains how personal information is collected, stored, shared and the plan should be accepted by the people.

4.1 Data Collection

The collection of personal data should obtain authorization of the law or the consent of people. The privacy protection laws vary between countries but generally, they will not hinder epidemic management measures. For example, GDPR (General Data Protection Regulation) states that personal data processing can be done without the consent of people if it is necessary for the public interest [10].

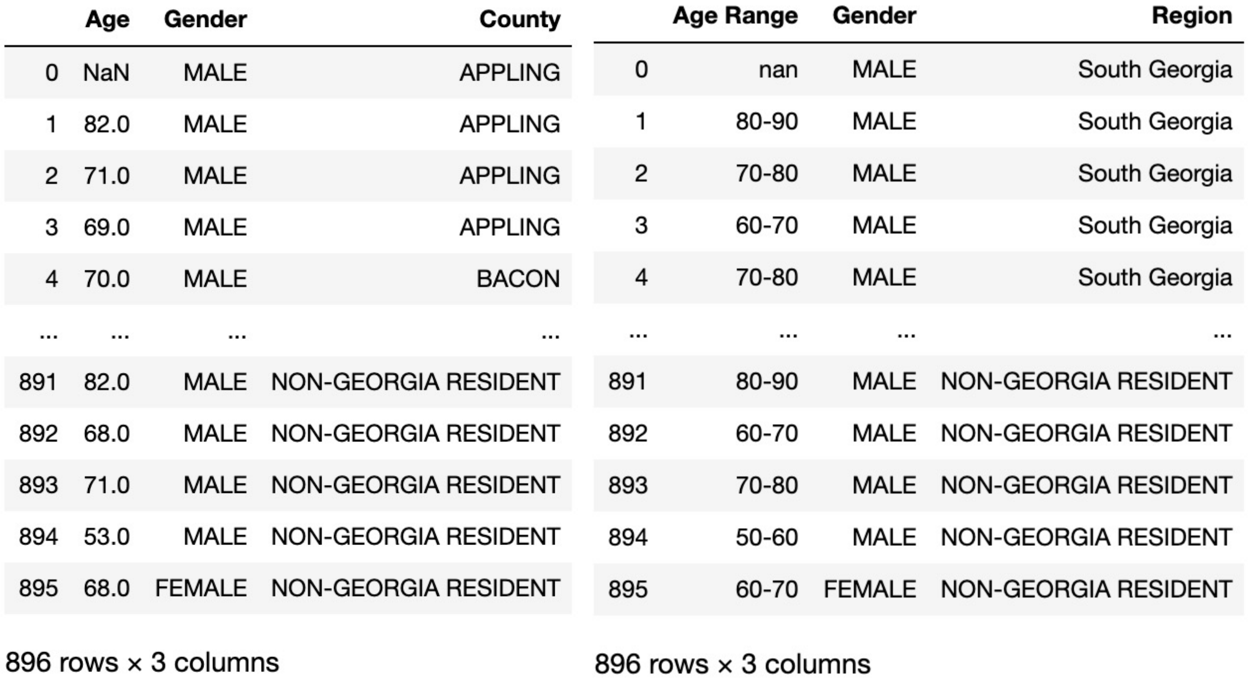
Even in emergency health situations, the basic privacy protection principles determined by law still need to be followed if possible [3]. Personal information should only be used for the purpose of epidemic control and follow the principle of data minimization. Public health experts need to specify necessary data to avoid the collection of irrelevant personal information.

Before the collection, the government should explain to people the purpose and benefits of the information collection, how long the data would be used, the possible scope of use, and the type of data needed to be collected. Transparency would help improve public trust in the medical sector and make future cooperation easier [4].

4.2 Data Sharing

For epidemic prevention, although the government needs to update and share the confirmed cases with the public in time, the protection of personal privacy should still be applied during the disclosure [2]. Openness and transparency are the keys to overcome panic. The publication of confirmed information guarantees the citizens' right to know and helps to raise the public's awareness of self-protection and enhance travel safety.

For the data sharing with the public, the government should only publish the information closely related to the spread and prevention of the virus, such as contact and traveling history of the confirmed patients. The names, mobile phones, and IDs of infected patients that can identify specific individuals should not be shared. Relevant departments should anonymize and desensitize the information during the information disclosure. The data could only be shared if personal identity cannot be recovered. For example, we could apply generalization to the dataset for Georgia death cases and convert the age to be a range and use region instead of the county. The original dataset would have around 60 percent of people be uniquely identified, and the dataset only have 24 people been identified after the changes.



	Age	Gender	County
0	NaN	MALE	APPLING
1	82.0	MALE	APPLING
2	71.0	MALE	APPLING
3	69.0	MALE	APPLING
4	70.0	MALE	BACON
...	...	...	...
891	82.0	MALE	NON-GEORGIA RESIDENT
892	68.0	MALE	NON-GEORGIA RESIDENT
893	71.0	MALE	NON-GEORGIA RESIDENT
894	53.0	MALE	NON-GEORGIA RESIDENT
895	68.0	FEMALE	NON-GEORGIA RESIDENT

896 rows × 3 columns

(a) 630 people could be uniquely identified

	Age Range	Gender	Region
0	nan	MALE	South Georgia
1	80-90	MALE	South Georgia
2	70-80	MALE	South Georgia
3	60-70	MALE	South Georgia
4	70-80	MALE	South Georgia
...	...	...	...
891	80-90	MALE	NON-GEORGIA RESIDENT
892	60-70	MALE	NON-GEORGIA RESIDENT
893	70-80	MALE	NON-GEORGIA RESIDENT
894	50-60	MALE	NON-GEORGIA RESIDENT
895	60-70	FEMALE	NON-GEORGIA RESIDENT

896 rows × 3 columns

(b) only have 24 people been identified

Figure 4: Applying Generalization techniques

4.3 Data Storage

Previously, some Asian countries have launched their own App applications, advocating ”centralized” data storage. The system is mainly based on the information reported by users, combined with their location information. The medical systems could use these data to determine the risk level centrally and better track the spread of viruses among people. However, the technology jointly developed by Apple and Google does not support the centralized storage of data due to the privacy concern. They recommend using a decentralized method to store information on personal mobile phones [11].

The decentralized storage could better protect user privacy but weaken the ability of public health institutions. It is probably better to allow the user to determine where data is stored. Personal information collected in the context of epidemic prevention is sensitive and should eventually be destroyed at the appropriate time after the end of this epidemic. The governments should be urged to formulate corresponding provisions to automatically terminate surveillance at specific times to prevent the normalization of emergency measures.

4.4 Conclusion

In general, the world is facing a significant challenge of COVID-19 and technology will be key to fight against it. The question now is not if privacy should be used but how they should be used. The government should take measures to minimize the negative effect on human rights when controlling the virus.

References

[1] A. Dukakis. China rolls out software surveillance for the covid-19 pandemic, alarming human rights advocates, April 14, 2020. URL https://abcnews.go.com/International/china-rolls-software-surveillance-covid-19-pandemic-alarming/story?id=70131355.

[2] M. Ienca and E. Vayena. On the responsible use of digital data to tackle the covid-19 pandemic. *Nature Medicine*, pages 1–2, 2020.

[3] P. L. Julie Brill. Preserving privacy while addressing covid-19, Apr 20, 2020. URL https://blogs.microsoft.com/on-the-issues/2020/04/20/privacy-covid-19-data-collection/.

[4] K. Klonowska and P. Bindt. The covid-19 pandemic: two waves of technological responses in the european.

[5] L. Lin. How coronavirus is eroding privacy, Apr 15, 2020. URL https://www.wsj.com/articles/coronavirus-paves-way-for-new-age-of-digital-surveillance-11586963028.

[6] B. McKenzie. Data privacy and seurity survey, April 17, 2020. URL https://www.bakermckenzie.com/-/media/files/insight/publications/2020/04/covid19-data-privacy--security-survey17-april.pdf.

[7] P. H. O'Neillarchive. How apple and google are tackling their covid privacy problem, April 14, 2020. URL https://www.technologyreview.com/2020/04/14/999472/how-apple-and-google-are-tackling-their-covid-privacy-problem/.

[8] R. Z. Paul Mozur and A. Krolik. In coronavirus fight, china gives citizens a color code, with red flags, March 1, 2020. URL https://www.nytimes.com/2020/03/01/business/china-coronavirus-surveillance.html.

[9] I. Tham. No other way but to make use of tracetogether mandatory, May 1, 2020. URL https://www.straitstimes.com/singapore/no-other-way-but-to-make-use-of-tracetogether-mandatory.

[10] R. Wang. The decision-making balance between "privacy" and "public health" ——10 consensus, differences and challenges in the protection of personal information in various countries, Mar 31, 2020. URL https://www.tisi.org/13613.

[11] Z. Whittaker and D. Etherington. Qa: Apple and google discuss their coronavirus tracing efforts, Apr 13, 2020. URL https://techcrunch.com/2020/04/13/apple-google-coronavirus-tracing/.