## Research Question

How does digital surveillance affect public political discussions on the Internet? Digital surveillance refers to the tracking, collection, and analysis of individual or group activities and communication on the Internet by governmental agencies or private enterprises. This surveillance may include, but is not limited to, email communications, social media behavior, online transactions, and tracking browsing histories. Due to such monitoring, the public may become more cautious and self-censored, especially when dealing with sensitive or controversial political issues. A notable example is the USA PATRIOT Act, which significantly restricted individuals' privacy by demanding broader surveillance and scrutiny of personal communications and online activities. Previous studies on this act have indicated widespread public concern and fear regarding the surveillance of privacy both online and in reality. Further research has pointed out that individuals who feel surveilled are more inclined to self-censor and are less willing to use the Internet for communication. Although prior research has discussed the impact of digital surveillance on public Internet usage preferences, it was mainly limited to measuring the issue through surveys or interviews. While surveys and interviews can reflect public preferences, they may not necessarily capture the behavioral changes of the public under the influence of digital surveillance.

The enactment of the Hong Kong National Security Law provides a new case study for using observational data to research the real-life impact of digital surveillance on public political discussions on the Internet. This law, claiming to uphold national security within the Hong Kong Special Administrative Region by combating acts such as secession, subversion of state power, terrorist activities, and collusion with foreign forces, led the Hong Kong government to prosecute a large number of dissenters and political activists who had participated in movements, using their statements on messaging apps or social media as evidence for charges. This implies that the law could trigger concerns among citizens about being prosecuted for their speech, thus fostering a sense of being under digital surveillance, potentially leading to self-censorship and a decreased willingness to engage in political discussions. To study the impact of digital surveillance on public political discussions on the Internet in the real world, I plan to collect all posts of political discussions from Hong Kong's largest online forum, LIHKG, and observe the impact of the enactment of the Hong Kong National Security Law and the resulting sense of being surveilled on public political discussions.

## Research Design

### Data

I plan to collect all posts since the inception of the website LIHKG, the largest online forum in Hong Kong, renowned for its political discussion boards. LIHKG played a pivotal role during the 2019 Hong Kong anti-extradition bill protests, serving as a crucial platform for the decentralized social movement. Previous research has indicated that a significant number of activists used it for coordination, mobilization, and information exchange. Concurrently, LIHKG hosts the most active online community in Hong Kong, drawing the largest user base for everyday discussions and sharing. This implies that utilizing the posts from this site as data will allow this project to amass the

most extensive and diverse collection of information on Hong Kong's political discourse.

**Variables**

I intend to examine the impact of digital surveillance (the independent variable) on political discussions on the Internet (the dependent variable). My approach will involve using the enactment of the National Security Law as a theoretical point for the emergence of a sense of digital surveillance and categorizing posts into discussions that are illegal under the law (such as secessionist and terrorist rhetoric), ordinary political discussions (daily criticism or complaints against the government), and discussions unrelated to politics. I plan to compare the frequency and nature of discourse among these three categories of posts before and after the implementation of the National Security Law. Discussions that are illegal under the law and those unrelated to politics will serve as control groups within the research design, aiming to control the influence of direct legal violations and the overall activity level on the website on the research hypothesis. Collecting all posts since the website's inception aims to control the seasonal variation in political discussions and provide more flexibility for optimizing subsequent research designs. Data collection and processing will be conducted through web scraping techniques and a large language model for data annotation. Subsequently, I initially planned to use time series analysis to examine the data.

**Feasibility**

**Collecting Method**

In the MACS30122 course, I developed a web scraper for LIHKG based on Selenium, enabling it to circumvent the site's anti-scraping mechanisms for dynamic data extraction. Over two weeks, I employed my personal laptop to scrape 50,000 posts. The website currently hosts approximately 4 million posts so that this method may require further optimization. I am contemplating leveraging the concepts of parallel computation with CPUs and GPUs discussed in the MACS30123 course to refine my scraping strategy and evaluate the feasibility of using the University of Chicago's computing cluster, Midway, for this purpose. As I aim to complete my writing sample during the summer but am uncertain about my access to the university's computing cluster for data collection during this period, I am also exploring commercial computing platforms as alternative solutions.

**Processing Method**

I plan to utilize ChatGPT 4.0 to annotate posts' types and discourse attributes (sentiment, stance). Previous research has indicated that ChatGPT can achieve an accuracy rate of approximately 70% for classifying social media data from platforms like Twitter and Reddit, allowing for the bulk annotation of large datasets. Through my preliminary testing, I found that after providing examples of convictions under the Hong Kong National Security Law related to speech, it could accurately assess the content of posts and classify them accordingly. However, given the vast number of posts and the cost associated with ChatGPT's API, I am still evaluating more economically viable

alternatives.
In the analysis part, I am still searching for an effective approach to analyze large volumes of observational data (social media posts). I have observed that many prior studies in similar contexts employ time series analysis or interrupted time series analysis to process data, but I am still learning this methodology. This semester, I have enrolled in a course on advanced causal inference offered by the Political Science Department. After further specifying the data and variables, I plan to consult with my instructor regarding feasible methodologies for this data.
.

**Evaluation of the fit**

I think the current strength of my research design lies in its effective capability to measure the dynamics of public political discourse changes, including both frequency and discourse. Moreover, this study possesses good potential for generalization by utilizing a comprehensive dataset from Hong Kong's largest political discussion website. However, a limitation in the current research design, as I perceive, is the necessity to integrate additional theoretical frameworks or research designs that directly link the enactment of the Hong Kong National Security Law with digital surveillance. Furthermore, there is a need for further optimization in the description of data processing and analysis methodologies.

**Possible advisor**

1. Dr. Molly Offer-Westort

Dr. Molly Offer-Westort is an Assistant Professor in the Department of Political Science and the instructor for my course on advanced causal inference this semester. She holds degrees in Statistics and Political Science and is a specialist in causal inference methodologies and political science research. Her rich knowledge and experience in causal inference for policy research could greatly assist me in refining the design of my data collection and analysis and optimizing the overall design of my policy study.

2. Dr. Ali Sanaei

Dr. Ali Sanaei is an Associate Instructional Professor in the MACSS program. He is also the instructor for my course MACS30100 and my preceptor. He has been trained in Political Science and has extensive experience collecting and analyzing public opinion data. He is also an expert in computational social science methods. I believe he can provide valuable advice and feedback on various aspects of my research.

3. Dr. Nick Feamster

Dr. Nick Feamster is a Professor in the Department of Computer Science. I enrolled in his course on Internet Censorship and Online Speech last quarter. I completed a project related to my current proposal (analyzing the sentiment changes of LIHKG users towards specific keywords before and after the National Security Law). He has a wealth of experience researching network security and censorship and focuses on using machine learning methods to gather and analyze information about human activities. He could offer significant help with the data collection and analysis methodology.